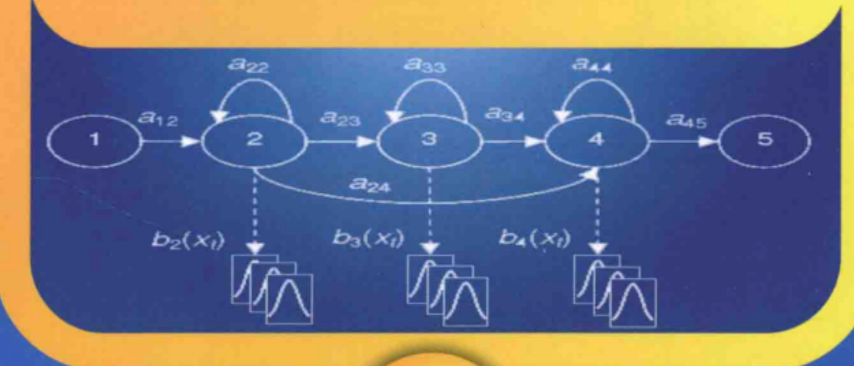
NEWTON QUANT 牛顿量化 | 量化投资方法丛书

刘振亚◎主编



解密复兴科技

基于隐蔽马尔科夫模型的时序分析方法



刘振亚
邓 磊
著

Decoding Renaissance Technology：

Time Series Analysis with Hidden Markov Model



中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE

刘振亚教授主编的量化投资方法丛书，一定会对中国基金业未来发展注入新的活力！

——中国证券投资基金业协会会长 孙杰

量化投资是现代资产管理的重要手段。随着中国证券和期货市场的逐步成熟，量化投资的理念和方法必将成为市场的主流之一。

——摩根大通期货有限公司董事长 周小雄

刘振亚教授从事量化投资研究和实践多年，终于将研究成果付梓，由衷地感到高兴!

——中国人民大学金融与证券研究所所长 吴晓求

上架建议 金融投资



体验更多精彩阅读
尽在中国经济出版社微信平台
请扫描二维码或查找zgjjcbs



ISBN 978-7-5136-3147-1

9 787513 631471 >

定价：30.00元

NEWTON QUANT 牛顿量化 | 量化投资方法丛书

刘振亚◎主编



解密复兴科技

基于隐蔽马尔科夫模型的时序分析方法

刘振亚
邓 磊
著

Decoding Renaissance Technology:

Time Series Analysis with Hidden Markov Model


中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE

图书在版编目（CIP）数据

解密复兴科技：基于隐蔽马尔科夫模型的时序分析方法／刘振亚，邓磊著.
北京：中国经济出版社，2014.4
ISBN 978－7－5136－3147－1
Ⅰ.①解…　Ⅱ.①刘…②邓…　Ⅲ.①对冲基金—金融公司—企业管理—研究—美国　Ⅳ.①F837.123
中国版本图书馆 CIP 数据核字（2014）第 049135 号

选题策划　张晓榕
责任编辑　李　博　李　珂
责任审读　霍宏涛
责任印制　张江虹
封面设计　任燕飞

出版发行　中国经济出版社
印 刷 者　三河市腾飞印务有限公司
经 销 者　各地新华书店
开　　本　787×1092　1/16
印　　张　11.75
字　　数　217 千字
版　　次　2014 年 4 月第 1 版
印　　次　2014 年 4 月第 1 次印刷
书　　号　978－7－5136－3147－1
定　　价　30.00 元

中国经济出版社　网址 www.economyph.com　社址 北京市西城区百万庄北街 3 号　邮编 100037
本版图书如存在印装质量问题，请与本社发行中心联系调换（联系电话：010－68319116）

总　序

量化投资与现代科技

量化投资方法被广泛应用在国际对冲基金中。在过去的三十年里，由于计算机技术和统计分析技术的进步，量化投资方法得到了迅猛的发展。

作为量化投资基金中的杰出代表，数学家西蒙斯（Jim Simons）所领导的复兴科技公司（Renaissance Technology）可谓独树一帜——他的大奖章基金在1988—2008年的20年时间里创造了年均收益35.6%的奇迹。1958年，20岁的西蒙斯从麻省理工大学数学专业本科毕业后，转入加州大学伯克利分校攻读数学博士，1961年博士毕业后回母校麻省理工大学任教。一年后，他跳槽到哈佛大学任教，又在1964年进入美国国防分析研究院工作。1967年，西蒙斯出任纽约大学石溪分校数学系系主任。在此期间，他与著名华裔数学家陈省身合作，创造了著名的陈-西蒙斯理论，并于1976年获得美国数学学会的威布伦奖。1978年，他离开石溪大学，成为职业投资人。1988年3月，西蒙斯成立复兴科技公司。

除了西蒙斯，复兴科技的三位元老Leonard Baum、Henry Laufer和James Ax都是一流的数学家，对复兴科技的长期发展产生了很大的影响——Baum是西蒙斯在国防分析研究院的同事，统计学中著名的Baum - Welsh算法的发明者，该算法被广泛应用于隐蔽的马尔科夫模型、语音识别、生物和应用统计中；James Ax也曾任石溪

大学数学系主任，在数论和几何学方面造诣颇深，曾在1967年获得美国数学学会数论方面的科伦奖；Henry Laufer也曾是普林斯顿大学数学教授，退休后在石溪大学创办了量化投资专业，专业课程包括：概率论与统计方法、线性规划、线性几何、数据分析、随机微积分、金融计量、最优化算法、资产定价、投资组合、金融市场、衍生品定价、固定收益产品等。

当然，成功的量化投资基金，除了数学家西蒙斯所领导的复兴科技，还有计算机教授肖尔（David Shaw）领导的肖尔公司，物理学家哈丁（David Harding）领导的元胜资本（Winton Capital），经济学家格里芬（Kenneth Griffin）领导的大本营投资（Citedal）等。

一些国际知名的对冲基金对人才的要求也可归纳为以下内容：很强的电脑编程能力（C，C++；API，FIX；R，Matlab，SAS等）；很强的数学或统计学分析技能（线性和非线性时序分析，数据挖掘，隐蔽马尔科夫模型，随机分析等）；很强的大型数据库处理能力；对衍生品、资产定价、市场微结构等有深入了解。

由此可以看出，量化投资方法是现代金融理论、现代数理方法和现代信息技术的综合体，所涉及的金融理论主要包括：资产定价，投资组合，衍生品定价，市场微结构，行为金融学等；信息技术主要包括：编程技术（C，C++），数据库技术，交易底层通讯技术（API/FIX）；数理方法主要包括：经济计量分析基础，线性和非线性时序分析，数据挖掘，隐蔽马尔科夫模型，随机分析等。

现代金融理论为量化投资提供了科学的理论基础。资产定价、衍生品定价、行为金融学和市场微结构是量化投资策略设计的科学基础；资产组合理论是量化投资降低和控制风险的有利工具。

现代信息技术和计算机技术为量化投资提供了坚实可靠的工具。软件工程和数据库技术是量化投资策略程序化的基础；API/FIX是交易底层自动化的基础。以上两者的结合，使得量化投资能够实现完全自动化交易。有人曾说，即使西蒙斯将其所用策略公众于世，能够把公式变成钱的人在全球范围内仍然是屈指可数。可见电脑技术和通讯技术在量化投资中的重要作用。

数理方法是量化投资最为核心的部分，也是数据分析和交易策略设计以及评估的基础。其所涉及的领域也是五花八门，仁者见仁、智者见智，主要包括：经济计量分析基础，线性和非线性时序分析，数据挖掘，隐蔽马尔科夫模型，随机分析，小波分析等。西蒙斯曾经说过，如果偏离了数学模型，将对我们公司没有任何好处。

在过去近十年的时间里，作者本人作为摩根大通期货公司（JP Morgan Futures Co.）的董事，有机会接触到一些国际知名的对冲基金，共同进行深入的研究和合作，元胜资本（Winton Capital）就是其中的一个。作为全球最大的管理期货（期货对冲基金），元胜资本管理着约300亿美金的资产，在其创始人科学家哈丁先生的领导下，在1987年至今这25年的时间里创造了16%—17%的年均收益。仅从年均收益的数字来看，元胜资本不如复兴科技可观，但大家要知道的重点是，复兴科技的大奖章基金的规模只有约50亿美元，而元胜资本的规模则接近300亿美元。

元胜资本现有员工200多人，近50%为研究人员，远离热闹非凡的一线业务，专心从事研究工作。若简单地按照国际对冲基金2%—20%的收费标准，元胜资本的员工创造了人均产出近亿元人民币的奇迹。可见，现代科技与金融的结合能够创造出巨大的财富，量化投资行业绝对是高科技产业中的佼佼者。

在过去的三十多年里，我一直从事计量经济方法和信息技术方面的学习和研究，本系列丛书也算是我和我的学生们对过去多年学习和研究的小结。这里，首先要感谢我的中国老师们：中国人民大学财金学院的黄达教授（我的博士导师，前校长），陈共教授（前财政系主任），王传伦教授；经济学院的杜厚文教授（前国际经济系主任，前副校长），王景新教授（我的硕士导师）；信息学院的陈禹教授（前院长），方美琪教授（我的本科导师），魏权龄教授，张怡兰教授，严颖教授；北京大学数学系的王萼芳教授。感谢我的美国老师们：George Horwich 教授（美国普渡大学，我的博士后导师），Roger Gordon 教授夫妇（加州大学圣地亚哥分校）和 Mark Machina 教授（加州大学圣地亚哥分校），Kajal Lahiri教授（纽约大学阿尔巴尼分校）和 Harold Watts 教授（哥伦比亚大学）。他们的教育使我领略了数学的美妙、计算机的精巧、计量经济的严谨以及金融市场的奥妙，为日后的量化投资研究奠定了

坚实的金融、经济、数理和计算机基础。

非常感谢中国人民大学校长陈雨露教授，金融与证券研究所所长吴晓求教授（校长助理），财金学院院长郭庆旺教授，副院长赵锡军教授，他们多年的友情、宽容和理解使得我能够有充分的自由和时间，在中国和英国两地从事我所感兴趣的研究项目。我也非常感谢伯明翰大学商学院的同事们：David Dickinson（前院长），Nicholas Horsewood，Robert Elliott，Toby Kendall，Alessandra Guariglia 和 William Pouliot。

感谢摩根大通期货公司董事长周小雄先生多年来兄长般无微不至的关照。感谢元胜资本的创始人 David Harding，元胜亚洲 CEO Charles Allard，Kurt Settle 和田野先生。与元胜资本的合作，尤其是元胜资本每年的年会使我眼界大开，受益匪浅。感谢 AHL－牛津量化金融研究院（AHL－Oxford Institute）使我有机会参加他们举办的世界一流水平有关量化投资的研讨会。

最后，感谢我在中国人民大学和英国伯明翰大学所指导的学生们，尤其是中国人民大学财金学院的博士生、硕士生和实验班的学生们，他们根据我的讲稿帮助整理了这套丛书中的很多内容。尤其是我的博士生杨武（现任教于中央财经大学），唐滔（现任职于中国人民银行研究所），罗涛（现任职于国家审计署），邓磊（现任职于北京工商大学），陈宇，张庆雪，刘琳，葛静，李伟；还有我在伯明翰大学的博士生曹璠玟，汪仕炫，张昆（阿斯顿大学），他们对量化投资的深入研究督促着我不断学习新的知识。这些学生们对新知识的追求，使得我多年来不敢懈怠，不断学习。从他们身上，我看到了中国量化投资业的未来。

出版量化投资方法丛书的主要目的是为国内的投资者系统地介绍有关量化投资理论、技术和方法，国际上成功的量化对冲基金公司，以及量化投资的研究成果和量化投资产业的发展动向。本丛书内容主要包括：经济计量分析基础，投资组合理论与实践，时序分析与神经网络，金融数据挖掘，解密复兴科技，随机分析，小波分析等。

我深切地希望，此套丛书能够为中国的证券、基金、期货、私募以及个人投资

者提高量化投资水平起到抛砖引玉的作用。

写这个总序前后花了我三、四个月的时间，从北京写到香港，从香港写到英国，从英国写到法国，最后又从法国写回北京。每次提笔都是千言万语涌上心头，本人最大的心愿就是在未来20到30年时间里，中国能够出现复兴科技和元胜资本这样世界一流的对冲基金。

为此，我愿奉献一生！

2013年9月于中国人民大学

前　言

马尔科夫模型广泛应用于信息通讯、计算机科学、生物遗传学、金融学、经济学等领域。出版本书的主要目的就是系统地介绍基于隐蔽马尔科夫模型（HMM）的时序分析方法及其在量化投资中的应用。

基于马尔科夫模型的时序分析方法的重要性，可以从复兴科技公司关键人物的研究背景中略见一斑，他们包括：Leonard Baum，著名的 Baum - Welsh 算法的创始人，该算法解决了不可观察变量概率的计算问题，被广泛应用于语音识别和信息解码；Elwyn Berlekamp，统计信息理论的专家；Nick Patterson，剑桥大学数学博士，国际顶级的隐蔽马尔科夫模型的专家，1993 年加入复兴科技公司；此外，还有原 IBM 实验室的语音识别专家 Peter Brown、Robert Mercer，以及该实验室进行机器翻译研究的其他专家。

人们可能要问，复兴科技为什么要雇佣世界上最优秀的语音识别专家和机器翻译专家呢？复兴科技研究人员给出的答案是："投资和语音识别，二者很相似，都是预测下一步将要发生的事情。"曾在 Google 工作过的腾讯公司副总裁吴军博士在其所写的《数学之美》一书中多次提到这些方法的重要性。2012 年，我曾买了几十本《数学之美》送给中国人民大学财金学院实验班的学生们。

由于传统的经济计量方法在预测精度方面存在极大缺陷，计量经济研究人员在经历了 20 世纪 60—70 年代的痛苦挣扎后，在 20 世纪 80 年代开始将注意力转移到对时序分析方法（Time Series Analysis）的研究。尤其是，1980 年初，C. Sims 在顶

级计量经济学杂志《Econometrica》上发表了著名的《Marcoecnomics and Reality》一文后，时序分析方法成为计量研究的主流。在时序分析方面，美国加州大学圣地亚哥分校（UCSD）经济系最为杰出，以 Cliver Granger、Robert Engle 和 James Hamilton 为代表人物。其中，Cliver Granger 和 Robert Engle 曾获 2003 年度诺贝尔经济学奖；UCSD 经济系主任 James Hamilton 所著《Time Series Analysis》一书自 1994 年出版以来，成为全世界顶级大学经济学、金融学博士生们研究时序分析方法的必读书目。

我之所以关注 UCSD 是因为 20 世纪 80 年代在福特班教我微观经济学和计量经济学的老师 Mark Machina 教授来自 UCSD。我从他身上学到了很多东西，尤其是他的讲课风格，对我影响极大。Mark 毕业于 MIT，是全球研究不确定性问题的大师。在不确定性领域，他对任何复杂问题都能深入浅出地讲解清楚。我从福特班毕业后，曾多次担任他的助教，受益匪浅。后来，在福特班教我宏观和微观经济学的教授 Roger Gordan 夫妇也到该系任教。

20 世纪 90 年代末至今，我陆续推荐自己的硕士生去 UCSD 攻读博士学位。尤其是前几年，我的博士生邓磊争取到了北京市留学基金会的奖学金，在我和肖志杰教授（Boston College）的推荐，以及 UCSD 孙一啸教授的帮助下，到 UCSD 进行了两个学期的学习访问。在这段时间里，邓磊学到很多东西、提高很快，我由衷地感到高兴。在此，对肖志杰教授、孙一啸教授的帮助和北京市留学基金会的资助表示感谢！

我最早接触马尔科夫链是在 20 世纪 80 年代初期，当时我在中国人民大学信息系学习运筹学。真正意识到它的重要性，还与芝加哥大学有关。1994 年，我在普渡大学进行博士后研究期间，我的博士后导师 George Horwich 带我去拜访他在芝加哥大学任教的几位老同学。我也顺便去看看我的学生、当时在芝加哥大学攻读博士学位的戴显峰，顺便关心一下他的学习情况。当小戴谈到计量经济课程时，提到了当时刚刚出版的《Time Series Analysis》一书。此时，我意识到了时序分析方法在未来的计量分析研究中的重要性。

2000 年初，我作为中国人民大学世界经济研究所所长主持召开了“世界经济与

中国”年会，并邀请了来自意大利、美国、英国和加拿大的专家学者们。在这次会议上，我结识了来自法国 Aix - Marseille 大学的 Eric Girardin 教授。Eric 毕业于英国剑桥大学，是应用计量方面的专家，他创建的全英文授课的金融计量硕士项目在法国教育部的专业排名中名列前茅。当他得知我在人民大学经济学院教授计量经济学时，提出合作研究的建议。我们最后决定将研究重点集中在如何利用 Hamilton（1989）的 MS - AR 模型来研究中国的股市。

此后的四、五年时间里，我每年都到普罗旺斯去拜访 Eric 一段时间，一边讲学，一边做研究；Eric 也每年都到北京，继续讨论我们的研究或修改我们合写的文章。在此期间，我们先后在 2003 年《Journal of Chinese Economic and Business Studies》的创刊号上发表了《The Chinese Stock Market：A Casino With “Buffer Zones”》一文；在《China Economic Review》2005 年第 4 期和 2007 年第 3 期发表了《Bank Credit and Seasonal Anomalies in China's Stock Markets》和《The Financial Integration of China：New Evidence on Temporally Aggregated Data for the A - share Market》等成果。

2007 年以后，由于我担任 JP Morgan Futures 的董事，Eric 担任亚洲开发银行学术委员会欧洲委员、香港金融管理局高级顾问，时间都很紧张，只能暂时放下我们的合作。

近年来，作为 JP Morgan Futures 的董事，我有机会接触到很多国际一流对冲基金的创始人、管理者和研究人员，参加他们的年会和研讨会。所有这些，都大大加深了我对量化投资方法的认识，尤其是意识到 HMM 和 MS - AR 模型在量化投资中的重要性。于是，在 2012—2013 年这两年的时间里，我为中国人民大学财金学院实验班的学生讲授了 Kalman Filter、HMM 和 MS - AR 模型及其在金融研究中的应用。本书的大部分内容都是根据我的讲稿整理、精简而成。

从表面上看，复兴科技这样的对冲基金公司与谷歌这样的高科技公司相差十万八千里；但从实质上来看，他们都是信息技术公司，都是依靠最新的科学理论、技术方法和最先进的计算机技术的高科技公司。他们的不同之处在于，谷歌公司研究的是如何从数以亿计的互联网数据中帮助人们找到最有用的信息，而复兴科技则是

从这些数据中找到能够判断金融市场上升或下降的信息。

本书分为四个部分：

第一部分主要介绍与此相关的数学基础：第一章介绍最大似然估计法，第二章介绍贝叶斯分析方法，第三章介绍马尔科夫链的基本知识。

第二部分将详细介绍 HMM：第四章主要介绍单变量的 HMM，而且状态也是基于齐次马尔科夫链，既没有趋势也没有季节变动；第五章和第六章将讨论 HMM 参数估计、预测与解码问题、隐蔽状态的估计问题、模型选择和模型检验等问题。

第三部分将主要介绍马尔科夫状态转换模型：第七章介绍序列不相关的马尔科夫状态转换模型；第八章介绍序列自相关的马尔科夫状态转换模型；第九章介绍 MS－AR 模型的估计方法。

第四部分提供了两个应用实例：第十章介绍 MS－AR 模型在宏观经济分析中的应用，第十一章介绍 HMM 和 SWARCH 模型在股市投资中的应用实例。

最后，感谢第一财经频道的陈琦、燕阳和上海对冲基金产业园邀请我为在沪的国内规模最大的几十家私募基金公司的高管们讲解有关复兴科技的方法和策略研究。感谢 Winton Capital（Asia）主席 Charles Allard 先生邀请我到 Winton Capital 英国总部做有关 HMM、MS－AR 模型和交易策略设计方面的演讲。

写于飞往阿姆斯特丹的飞机上

修改于英国伯明翰大学 J. G. Smith 楼

2013 年 12 月 31 日

如果偏离了数学模型，将对我们公司没有任何好处。

——复兴科技公司创始人　詹姆斯·西蒙斯

目　录

第一部分　基础知识

第二部分 隐蔽马尔科夫模型

第三部分 马尔科夫状态转换模型

第四部分　HMM 和 MS－AR 模型应用



引 言

解密复兴科技

作为量化投资基金中的杰出代表，“量化之王”数学家西蒙斯（Jim Simons）所领导的复兴科技公司（Renaissance Technology Corp.，简称“复兴科技”）可谓独树一帜，旗下规模为50亿美金的大奖章基金（Medallion）在1988—2008年的20年时间里，创造了年均收益超过35%的奇迹，这还要扣除5%的资产管理费以及44%的投资收益分成等费用，并经过严格的财务审计。

不仅于此，历史上的大奖章基金面对多次金融危机和政策波动都有杰出的表现。1994年，美联储连续6次加息，大奖章基金净赚了71%；2000年科技股股灾，标普指数下跌了10%，大奖章基金却大获丰收，净回报率高达98.5%；2008年，全球金融危机，各类资产价格下滑，大部分对冲基金都亏损，而大奖章基金净赚了80%。

美国著名对冲基金观察家Antoine Bernheim曾说过：“西蒙斯才是真正的NO.1，他超越了乔治·索罗斯（George Soros），马克·金顿（Mark Kingdon），布鲁斯·科弗纳（Bruce Kovner）和蒙罗·特劳特（Monroe Trout）。”

本书题目之所以不是揭秘，而是解密，是因为复兴科技不是仅仅使用一种方法，而是融合了很多种方法，我们在这里谈的是复兴科技最主要的方法之一：隐蔽马尔科夫模型（HMM）。为什么我们会认为HMM是复兴科技采用的主要方法呢？因为复兴科技主要核心人员的研究背景都与此有关。

第一节 西蒙斯与复兴科技

复兴科技是一家很杰出的基金公司，代表人物就是西蒙斯，他是一位优秀的数学家，优秀在哪里呢？他在20年的时间里创造了年收益率36%的奇迹，这是一件非常了不起的事情。国内凡是搞投资的人都在聊巴菲特，而众所周知，巴菲特的年收益率是26%左右。

西蒙斯出生在20世纪30年代，20岁的时候从MIT数学系毕业，后来到伯克利读了数学博士，1961年毕业后回到MIT任教，待了一年后又跳槽到哈佛任教。可能因为父亲是商人的缘故，他也非常具有企业家精神。1964年的时候，他进入美国国防研究院任职。这个新单位听起来就很神秘，西蒙斯在那里主要研究什么呢？就是研究怎么破解敌方的密码，怎么来编码使别人破译不出来。

1966年底到1967年初，西蒙斯发表了坚决反对越战的言论。当记者采访他对越战怎么看时，他说就不应该打这个仗，打这个仗就是错的。当时越战刚开始，这篇访谈一经刊登，顿时引起了轩然大波，是因为报道中特别强调了军中有人反对越战。当国防部长知道是西蒙斯说的之后，便毫不犹豫地把他解雇了。“我被解雇的时候感觉自己特别无力，”他说，“我当时就想，如果你是老板的话就没人能解雇你了。”

1967年，西蒙斯应邀到美国大学石溪分校，也就是著名物理学家杨振宁先生曾工作过的学校，做数学系的系主任。当时他才30多岁，非常年轻，在此期间他潜心于数学研究。当时该校还有一位很有名的数学教授，就是后来回到南开大学的陈省身教授，西蒙斯与他一起发现了数学理论里著名的陈－西蒙斯理论。

陈—西蒙斯理论对于投资领域的人来说或许有些陌生，但是该理论对其他学科产生了巨大的影响。在20世纪80年代中期，普林斯顿大学教授Edward Witten发现了该理论在物理学方面的适用性，并称之为陈－西蒙斯场论。现在陈－西蒙

斯理论已经作为一种重要的工具广泛应用于物理学研究的很多方面，包括弦理论和超引力黑洞的研究。一位从普林斯顿大学跳槽到麦肯锡顾问公司的数学家 Dennis McLaughlin 说："物理学家们每天都能依靠陈 - 西蒙斯理论发现新的研究方向。"

此后的十年里，西蒙斯获了不少的奖，其中最高荣誉是 1976 年美国数学学会的威布伦奖。

尽管他在数学研究领域里成绩突出，但具有创业家精神的西蒙斯很快厌倦了单调的科研生活。1978 年，西蒙斯离开石溪大学成为职业投资人。在真正成立复兴科技的 1988 年之前，他也办过一些实业：1961 年他曾和麻省理工的同学投资过一个哥伦比亚地砖和管道公司；在伯克利任教的时候，他曾投资 5000 美元去做婚礼礼物的生意。这些或许成功或许失败的投资经历最终使他转到了证券行业的投资领域。

第二节　复兴科技的元老们

除了西蒙斯之外，复兴科技最早的几位元老都是数学家。

一位元老是 Leonard Baum，他是一位很优秀的数学家，曾在复兴科技参与过模型研究。Baum 是西蒙斯在国防分析研究院的同事，Baum - Welsh 算法的发明者之一，该算法主要是用来解决不可观察变量的最大似然函数计算的问题——也就是说，在丢失了一些观察值或者变量是不可观察的情况下，应该怎么来处理。这个算法，在我们后面谈到隐蔽马尔科夫链的时候还会谈到，它在语音识别、生物和应用统计中也是很重要的。

第二位元老是 James Ax。复兴科技公司的前身是 Axcom 公司，后来才叫现在的复兴科技公司，是 Ax 创建的。西蒙斯与他两个人合伙运营的时候，还叫 Axcom。这位 Ax 是很优秀的数学家，1967 年的时候也获得美国数学学会在数论方面的奖，个

性非常强，结果造成两个人合不来。

最后一位是 Henry Laufer，他也是非常优秀的数学家，曾任普林斯顿大学的数学教授，并在复兴科技中一直担任首席研究专家的职务。从复兴科技退休后，他在石溪大学创办了量化投资专业。

第三节　复兴科技中的主要研究方法

隐蔽马尔科夫模型（以下简称“HMM”）对复兴科技很重要，因为从复兴科技主要人员的研究背景来看，他们最早应该是用这个模型的。当然，他们自己从来不会承认用的是哪个模型，这一点是人们对复兴科技公司的共识。

为什么这样说?

这先要从复兴科技的关键人物谈起。首先，鲍尔曼（Leonard Baum）是著名的 Baum - Welsh 算法的创始人，复兴科技的核心创始人之一，前面已谈过；另外一位是伯乐卡普（Elwyn Berlekamp），复兴科技最初的灵魂人物，也是数学教授，是统计信息方面的专家。他曾在 Ax 和西蒙斯不合的时候把复兴科技全部买了下来，当然包括 Ax 和西蒙斯两个人的全部股份。一年以后，公司业务又走上正轨，伯乐卡普将公司又出让给了西蒙斯。很多人说全世界最傻的人是他，但是，他说自己很开心，乐意做研究。他现在还在做教授，继续从事着研究工作。

另外一个重要事件是在 1993 年，复兴科技花重金把全世界顶级的 HMM 专家——剑桥大学数学博士帕特森（Nick Patterson）聘请来公司工作。如果复兴科技不用 HMM 的话，那么根本没必要这么做。

从算法上来说，HMM 算法可以用在语音识别和机器翻译上，也可以用在股市投资上。因为语音识别和机器翻译都是一个顺序问题，是编码的问题，股市波动也是一个升降的序列问题。

此外，复兴科技还把 IBM 公司 Watson 实验室的语音识别专家布朗（Peter

Brown)、麦瑟(Robert Mercer),还有该实验室进行机器翻译研究的主要专家全部挖过来了。

第四节 什么是 HMM?

下面,我们谈谈这个 HMM 到底是怎么回事。

众所周知,股价是可以观察到的,但是有一个大家平时不太注意的问题:相同的股价在不同的状态下有着不同的意义。比如说 3000 点对于现在(2100 点左右)来说就是高的,但是 3000 点在一个快速上升的牛市里面可能又是低的,或者是中等的。所以,我们不能光看这个数字,而要看其所处的状态,这两个东西要结合在一起看。但是,实际问题是股价所包含的状态是观察不到的。我们用什么办法能够把它从股价中提炼出来呢?这个问题很关键。

大家都清楚什么是股价,因为它是可以观察的,那么有人可能就会问,到底什么是状态?简单地说,股市有两种状态:一种是牛市,一种是熊市。但是,这两种状态是不可观察的,这就需要用可观察到的数据去估计这个状态。问题的复杂之处就在这里:有一些不可观察的变量,或者说有一些丢失的变量,我们怎么通过可观察的变量得到这些不可观察的变量值,并据此来判断我们所看到股价的真正意义。

由于股市的状态是不可观察的,所以我们一定要对这个状态怎么变化做出一些假设。否则,不可观察又不做一些假设,那真是摸不见看不着了。该怎么假设呢?这就需要用到转换矩阵。换句话说,我们先做以下的假设:如果熊市到熊市的转移概率,即今天是熊市明天还是熊市的概率是 0.8,那么熊市到牛市的概率就是 0.2;如果牛市到牛市的概率是 0.9,那么就只有 0.1 的概率明天是熊市。就是因为这个状态变量我们观察不到,所以我们要假设它服从这样一个转换矩阵。

HMM 三要素：

1. 可观察的状态依赖变量（股价）：p_1，p_2，……，p_t

2. 不可观察状态变量（熊牛）：S_1，S_2，……，S_t

3. 状态转换矩阵：

	S_t = 熊	S_t = 牛
S_{t-1} = 熊	0.9	0.1
S_{t-1} = 牛	0.2	0.8

当然，转移概率矩阵中的概率是 0.9 还是 0.1，是 0.8 还是 0.2，要通过可观察的变量来估计这些值。前面说过，不可观察变量是需要用可观察变量来估计的。同样，这个转移概率的值也是需要从可观察变量来估计的。要通过什么方法把这些值估计出来呢？Baum – Walsh 很好地解决了这个问题，主要是用 Expectation – Maximization，简称 EM 算法。

因此，HMM 主要的思想是这样的：人们所看到的股价数据隐含了不可观察的状态，这个状态需要我们首先对它进行相应的假设，然后再估计出来。

第五节　HMM 举例

下面举例说明观察值和状态的关系。假设 p_1 是某股票在熊市的一个收益分布，也就是说，熊市的时候股价在 13 元左右，它服从 p_1 这个分布；p_2 是该股票在牛市的一个收益分布，也就是说，牛市的时候股价在 28 元左右，它就服从 p_2 这个分布。可以简写为：

$p_1(x)$：某股票在熊市的收益分布；

$p_2(x)$：某股票在牛市的收益分布。

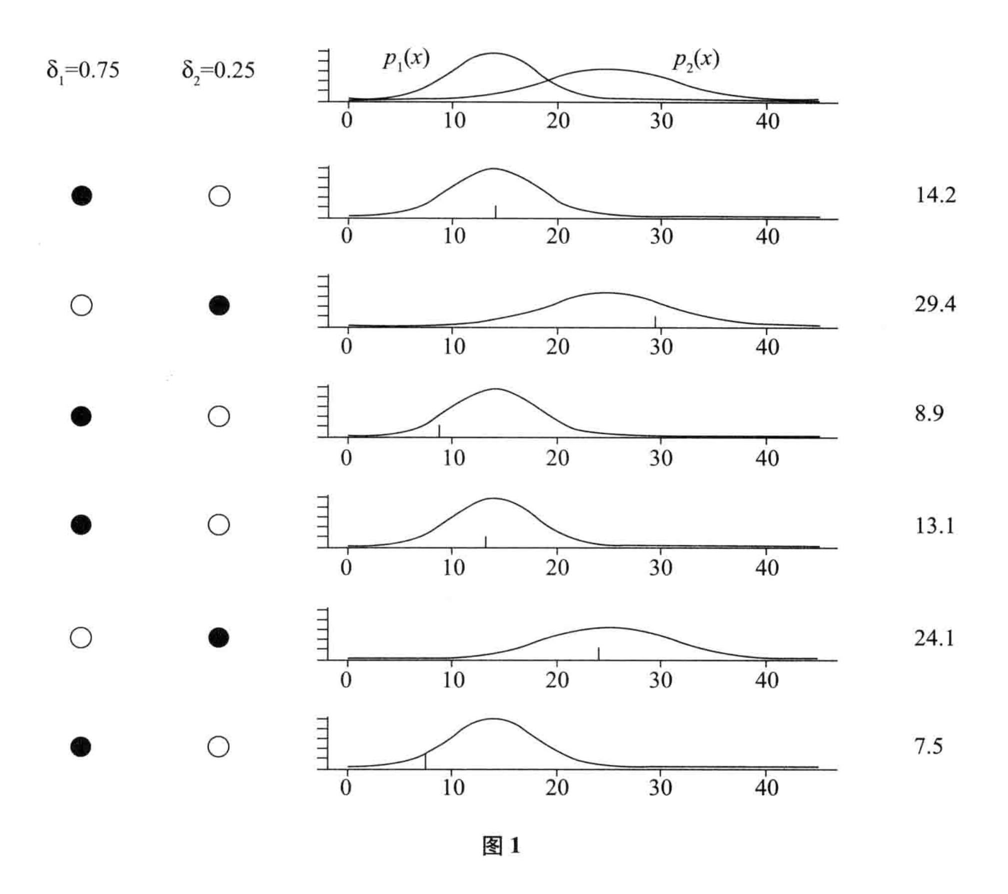


图 1

具体观察数据如图 1。第一个观察到的数据是 14. 2 元，请读者来判断一下这个数据是从哪一个分布产生的？我们可以认为，第一个观察到的数据 14. 2 元最可能是 p_1 分布（熊市）产生的。因为相对于 p_2 分布（牛市）分布而言，p_1 分布（熊市）可能性更大。也就是说，这个数据肯定是从 p_1 和 p_2 两个分布里面中的一个所产生的，但在这两个分布里面，14. 2 元更靠近哪个呢？答案肯定是靠近这个 p_1 分布。换句话说，14. 2 元所对应的状态最为可能是熊市。

第二个可观察到的数据 29. 4 元更靠近 p_2 分布。所以，看到第一股价数据 14. 2 元的时候，就会觉得这可能处在熊市的状态；看到第二个股价数据 29. 4 元的时候，人们可能认为是处在牛市的状态。

当观察到第三个股价数据 8. 9 元的时候，大家可以看看处于这两个分布中的哪

一个分布？可能是 p_1。如果大家看到第四个可观察的股价数据 13.1 元的时候，也可以很快就肯定是 p_1 分布所产生的（熊市状态），因为它更可能发生。那么，当看到第五个可观察数据 24.1 元的时候，这个数据很大可能不是 p_1 分布所产生的，说明当时状态是熊市的概率很小；而是从 p_2 分布中产生的概率是很大的，说明当时状态是牛市的概率很大。

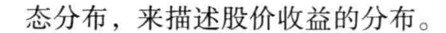
因此，大家可以看出，平时人们所观察到的价格，不仅仅是简单的价格，同时在它的背后有还有一个状态在决定着它。这一点很重要，千万不要认为股价在什么时候都是一样的。

一碗饭可能在吃饱的时候对于你来说是多了；但是当在饥饿的时候，你就会觉得一碗根本不够。同样一碗饭在不同状态下其价值完全不一样，股价也是同样，处在不同状态下意义是完全不一样的，这就是 HMM 需要解决的问题。

第六节 股价收益分布与 HMM

如果上面的例子中 p_1、p_2 都是正态分布，当然也可以假设它们是很多种其他类型的分布，还可以用非参的办法来解决，这些都没有问题。为了简单起见，我们这里假设其是正态分布。那么，这时股价收益的分布到底是什么样的情况，是不是符合正态分布呢？实际上股价的分布不服从正态分布，为什么呢？从图 2 中可以看出，正态分布是一个铃铛，一个标准的正态分布中 95% 的概率是在正负 1.96 之间。标准正态分布，－1 到 ＋1 之间出现的概率是将近 2/3 的；从－1.96 到＋1.96 出现的概率是 95%。

但是，股价收益实际分布不是像标准正态分布这么好，它两边出来了两个尾巴，即所谓的“肥尾”，如图 3 所示。若按照标准正态分布，在左右两边出现这样的概率几乎是不可能的。这就需要我们用三个分布的混合分布，而不是用一个简单的正



态分布，来描述股价收益的分布。

图 2

图 3

从图 3 中我们知道，股价收益的实际分布会出现“肥尾”——左右两边会翘起来。那么，这种情况该如何处理呢？该如何描述这样一个混合分布呢？此时，我们就需要用到三个分布，中间一个分布，两边各一个分布，来描述股价收益的实际分布情况。左边的分布是熊市；中间的分布是盘整；右边的分布是牛市。这样，用三个分布组成混合分布的转移概率矩阵是一个 3 阶矩阵，与前面的 2 阶转移矩阵一样，

其具体形式如下：

表 1

	熊市	盘整	牛市
熊市	0. 9	0. 05	0. 05
盘整	0. 1	0. 8	0. 1
牛市	0. 05	0. 05	0. 9

这时就出现了一个非常重要的问题：我们怎么把这样一个股价收益的分布拆成三个分布？这里使用的拆分办法就是用 Baum - Welsh 算法。使用该算法后，我们就可以知道这三个分布的均值是多少，方差是多少；拆成的三个分布有多少比例可能分配在中间这个分布，有多少比例是分配在右边这个分布，有多少比例是分配在左边这个分布；以及三个分布之间的转移概率矩阵。

一般来说，这三个有可能都是正态分布；也有可能都是是泊松分布；也有可能中间是正态，两边是泊松。根据数值算法或 EM 算法，我们可以得到这三个分布的均值和方差，而且还能够得到各个分布在这个混合分布的比重；更重要的事情是，我们还能够得到转移概率矩阵的估计值。

根据这些转移矩阵的估计值，我们可以算出处在这三种分布的期望持续期。这些期望持续期对于我们设计策略来说非常重要。换句话说，假如只允许做多，那么，我们能够赚钱的部分只有右面这个分布。如果右面这个分布的持续期是四期，我们要花费一期去观察它，因为只有在一期过后，我们才能够确认它处在这个分布。当确认某股票一进入右面这种状态，马上就买。为什么呢？因为它还会在这个状态持续三期。在第四期到来的时候，不管状态变不变，我们就要减仓。所以，每个状态的持续期很重要，它是策略设计中很关键的要素。

大家在 2006 年股市上升的时候很开心，主要是因为右边这个分布；而大家在 2008 年股市大跌的时候很难受，主要是因为左边的这个分布。如果了解了上述分布的期望持续期，那么我们将会知道：处在这两个分布的时间大概有多长；一旦进入到这个状态里面以后，还能待多久；该怎么办，是加仓还是减仓？

第七节 HMM 与交易策略设计

实际上，人们看到的只是股价和指数点位。但是，对于 HMM 来说，看到的不仅仅是这样的数据，还可以从股价中分离出不同的状态：是处在熊市状态还是牛市状态；是处在状态中的第几个周期。假如状态持续期是四期，那么，现在是处在第一期、第二期、第三期，还是第四期？这些信息对于我们设计投资的策略太重要了。

根据 HMM 设计策略时，会遇到很多问题。比如说，裁出来的三个分布都很接近，均值也很接近，这时候该怎么办？为了解决稳定性问题，可以取三个的中间值，采取所谓的均值回归的办法，过正 95% 的临界点就卖，它肯定会往中间分布。

上面谈的是第一个问题。第二个问题，上一节举的四个期的例子无论如何都要浪费掉一期去观察所处状态，那么如果我们算出来的持续期只有 1.6 怎么办？持续期不到两期，这个问题该如何处理？这是一个很头疼的问题。

第三个问题是持续期取决于转移概率矩阵，该矩阵的平稳性该如何检验？如何检验这些估计参数到底是不是稳定的，如果转移概率估计量不稳定，这个策略的设计也是很麻烦的。

再一个问题就是当数据量大的时候 HMM 的参数估计计算时间长短及其时效性，这一点也是很重要的。如果计算时间很长，做高频交易就会有问题，所以需要权衡。复兴科技是基于分钟或者秒的数据在做交易，为什么呢？因为 EM 算法有很重要的一个特点，如果说我们都用低频的数据，例如周数据或者月数据，这样裁出来的三个分布就很容易很接近，整体混合分布也比较容易接近于一个非稳态的正态分布；如果我们用秒的，或者是分钟的，拆出来的三个分布会离的很远，很容易判断，收益的机会就多；而且，这三个分布不仅离得很远，每个分布的持续期也会比较固定，这可能是另一个原因。

实际上，我们用中国股市的周数据计算出来的持续期就是1.6。那么，根据这个特点怎么来设计策略呢？这样的持续周期实在太短了，只能采用接近于 Larry Williams 的办法：先埋伏好，等突破了就自动买入。

第八节　基于 HMM 的交易策略

本节设计一个基于上证指数的一个非常简单的交易策略，数据来自2006—2013年每周的上证指数。读者在学习这个策略的时候一定要注意，不要简单地死搬硬套到实际中，否则肯定会出大问题。基于上证指数的这个策略是什么样的呢？如果现在的股价大于两个月的均价再加上这两个月波动的0.8倍，就买入。换句话说，这是一个突破就买入的办法，这里用的突破点是两个月的均价加上两个月的标准差的0.8倍。

图4是基于上证指数2006年1月至2013年11月的每周收益和上述策略的运行结果。

有的人会说，这个策略不好。为什么不好呢？图中大盘最高点已经到4.5倍这里了，从2006年1000多点到2007年的6000多点增长了4.5倍；而采用这个策略才赚到2倍。

的确，在2006—2008年这个策略看似不好，明显收益比大盘低。但是，在大盘2008年快速下跌的时候，投资者们就体会到了这个策略的好处。这个简单的策略没像上证指数给投资者们这么大的惊喜，同样，也没有给他们这么大的失望。2008年后这个策略的好处就凸显出来了，波动比上证指数小，在后面股市的一波反弹中，这个策略也赚到了钱。

尤其从2010年以后，这个策略一直好于上证指数的表现。

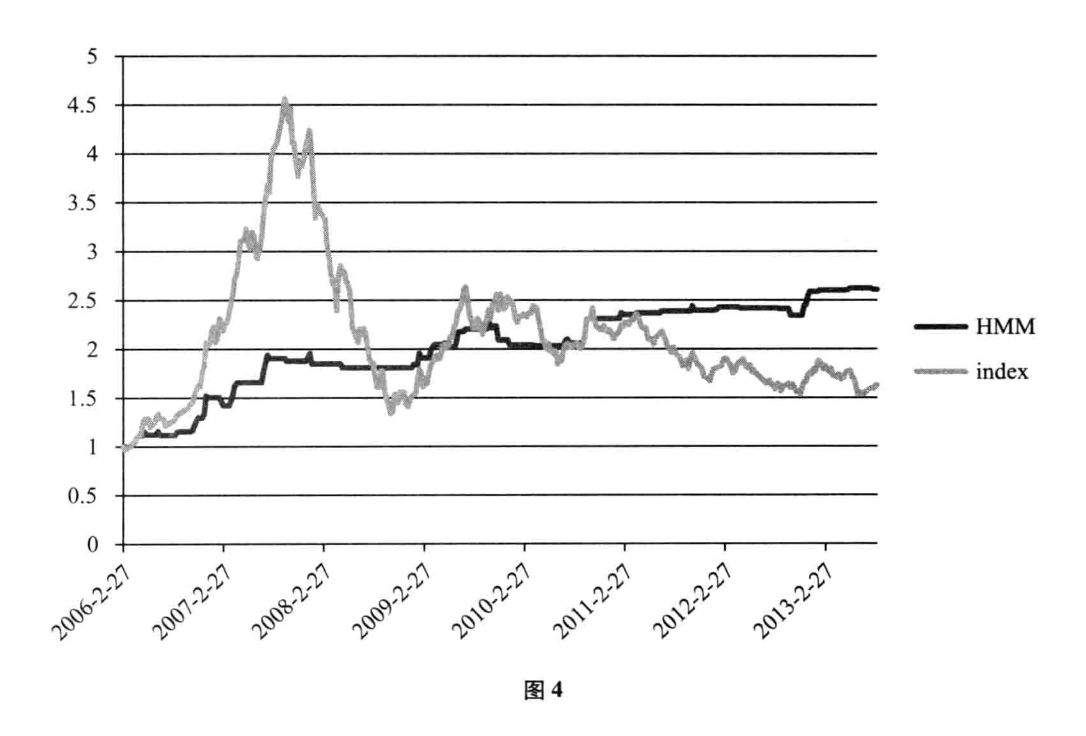


图 4

假设允许做空，这个策略会是怎样的呢？改变是必需的，因为做空的速度快一点。做多和做空二者结合，收益肯定要比只做多这个结果要好一些。

第九节 交易策略的评价问题

大家一定要记住，评价一个交易策略好坏很关键、很重要的问题就是，千万不要只看回报率，一定要看三个数据：

第一个数据当然是回报率，它是很重要的。

第二个非常重要的数据是 Sharpe - ratio。众所周知，Sharpe - ratio 是收益除以风险的比值，换句话说，它表示冒一分钱的风险能够得到多大的收益。这个比值是最最重要的，希望大家在具体应用的时候一定记住。如果 Sharpe - ratio 低于 0.8、0.9，这样的策略是不能用来管理大资金的，否则会带来很多很多的麻烦，比如说最

大回撤（MDD）。上一节提到的 HMM 策略回报率是 2.6，并没有比上证指数回报率多多少，尤其是这里还没有考虑到交易成本。当然每周做一次交易成本不一定很高，但如果做高频的话，交易成本就会很大。上证指数的 Sharpe - ratio 是 0.3，这个交易策略的是 1.2，二者差三倍多一点，说明这个策略明显好于上证指数。

第三个重要的数据就是最大回撤（MDD）。表 2 是上证指数和这个简单的 HMM 策略的结果对比：

表 2　上证指数和 HMM 策略的结果对比

	HMM 策略	上证指数
回报率	2.607144	1.622555
Sharpe - ratio	1.19993	0.367541926
MDD	-0.26956（10%）	-3.22777（70%）

这三个数据当中，MDD 是最重要的。HMM 交易策略只有 10% 的 MDD，而上证指数则达到了 70%，二者相差 7 倍。所以，换句话说，如果说你能够忍受和上证指数同样比例的回撤，那么你可以把交易杠杆放大七倍，现在的股指期货就可以很容易地做到这一点。

交易杠杆放大七倍后挣不挣钱呢？当然挣钱。图 5 就是放大七倍的结果，这是很吓人的。如果放大七倍，这个策略可以赚到一百多倍。但是，为什么我们不采取这个策略呢？读者们可以想象一下，开始时的一个亿，很快可以赚到 100 个亿，这是非常令人激动的。但是，这 100 个亿可能在某个时间点会亏掉 60 多个亿，这种亏损是否是投资者们能够承受得了的？如果说能够承受，那么这个策略是很好的。

国外机构评价一个交易策略的好坏时，其他的不要看，只看 MDD 和 Sharpe - ratio 这两个最主要的指标数据。在美国，学投资主要讲风险，不讲回报。可见，风险在金融研究和实践中的地位有多么重要！

对于这个 60% 多的大回撤，出现的第一个问题是，没有多少投资者的心理素质能够好到承受得了。第二个问题是，一旦有回撤，投资者就会把钱撤出基金，这样整个策略就无法继续下去。所以，基金经理一定要跟客户说清楚自己策略的历史最

大回撤是多少，问客户能不能承受这种回撤。如果客户承受不了，就可以把杠杆降低一下，杠杆放大后的回撤程度是很大的。

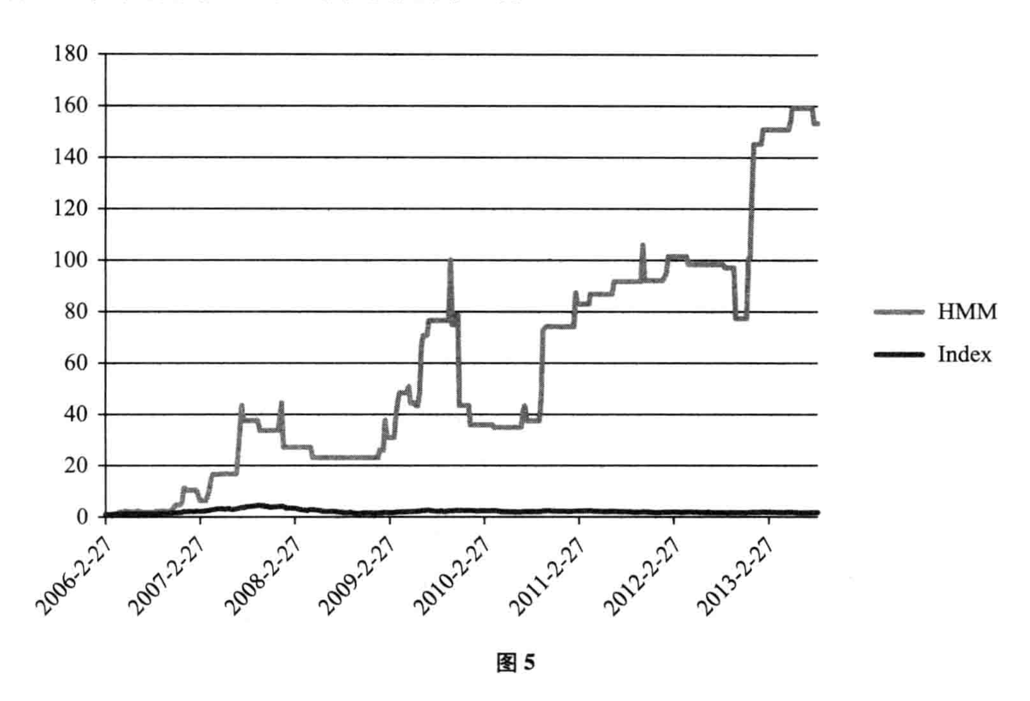


图 5

大家都知道，基金行业最需要的就是坚持，能坚持下来成为“长青树”是很不容易的。通过杠杆，盈利可以很容易从 2 倍到 100 多倍，但是问题是，面对可能出现的这种 60% 多的回撤，无论对基金经理自己，还是作为投资者的客户，在心理上都是不可能接受的。

下面，我们再比较一下 Sharpe - ratio。HMM 策略和上证指数二者只差了 3. 26 倍，我们就将杠杆设定为 3. 26 倍。在这种情况下，MDD 就容易接受得多，而它的收益是 16 倍。所以，今后大家一定不要只谈回报率，这没有任何的意义，重要的是 Sharpe - ratio 和 MDD。图 6 是这个简单做多的 HMM 策略放大 3. 26 倍的结果。

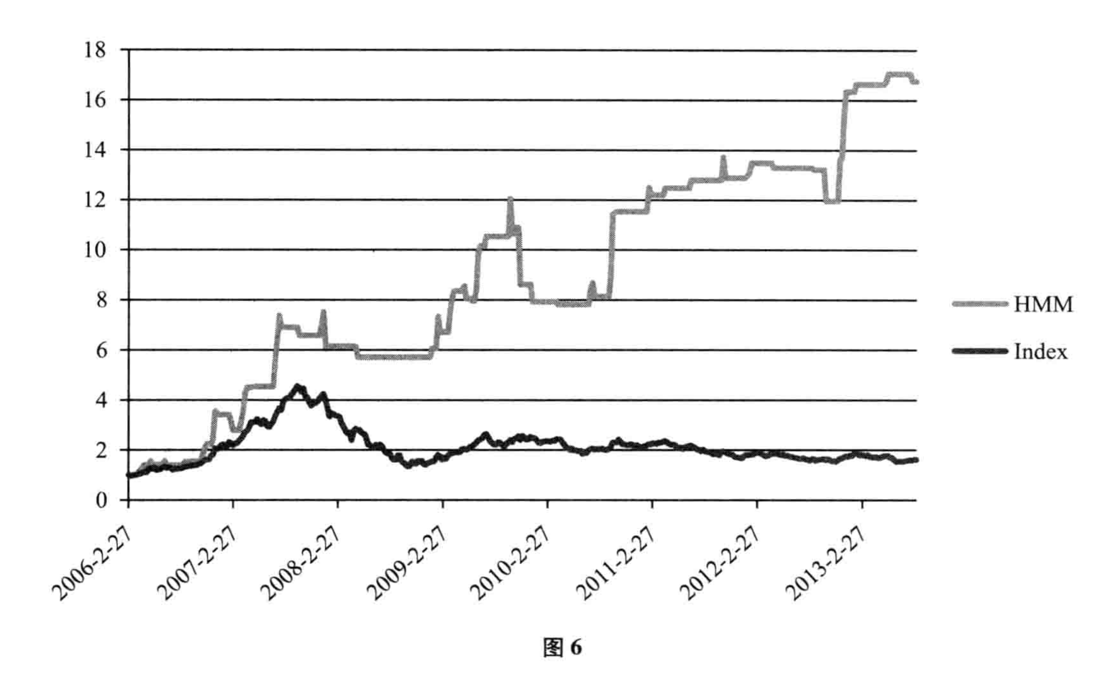


图 6

以上策略只是基于 HMM 设计的一个简单交易策略，大家可以在这个基础上做得更加深入一点。

另外要强调的是，Sharpe－ratio 也有它的问题，这牵涉到风险衡量。一般而言，Sharpe－ratio 用方差来表示风险。所以，要使风险最小，就必须使得整个方差缩小，这样同时把正向和负向的两个肥尾都往里靠。但是，如果我们只做多的话，就会是希望正向的肥尾越大越好，负的这边越小越好。因此，有人就提出来单边风险问题，例如，α－风险和一致性风险。

我们这里的数据用的是周数据。当然，可以用分钟、秒的数据；考虑更多的因素，交易成本和价格冲击问题的风险。高频数据拆开所得到的转移矩阵稳定性会更好一点。这样，策略结果也会更好一点，就是这样的道理。很多人都说，复兴科技实际上每天做数千笔的交易，前一秒钟上涨、下一秒钟上涨的概率比明天上涨概率的稳定性可能会更高。

第十节 科技与投资

上面只是举了一个简单的基于 HMM 设计交易策略的例子，大家就会发现 HMM 的计算量很大。从这里我们可以看出，投资是科学和技术的结晶，是真正的高科技。

量化投资主要包含三个方面：第一是理论，包括金融理论、经济理论，还有其他方面的一些理论；第二就是方法，比如说像金融计量方法、计算金融的方法、统计方法、单尾风险的衡量方法、小波分析、HMM 等；最后一个就是技术，有人说，即使西蒙斯告诉你算法，真正能够在技术上实现的也没有几个。量化投资的技术牵涉到哪些？常见的有 API，还有 FIX，这些接口怎么做；Database 怎么设计；研究的时候需要用到 Matlab，R 等等。

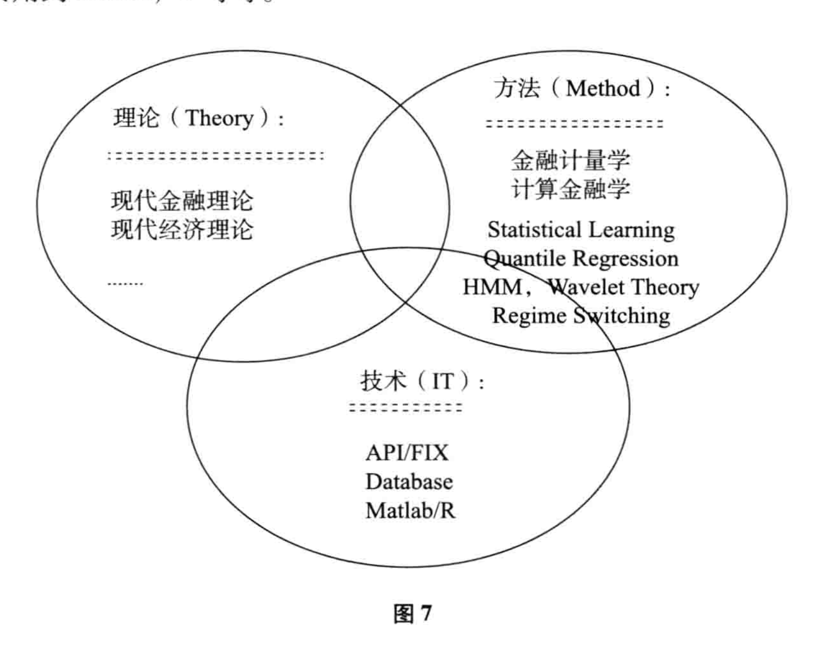


图 7

除技术之外，还要找到比较好的理论和方法，这三者的有机结合，才是真正的量化投资。专做理论的人可能认为数学是量化；专做技术的人认为计算机化是量化；

其实，以上的说法都不对，真正的量化是上面三个方面的综合。

同时，真正的量化投资应该尽量减少人工干预，全部由计算机来做。在100次的交易中，人工干预最多只有3到5次。真正的量化投资过程是研究人员根据历史的数据和科学的方法，找到一些规律性的东西，形成相应的投资方案，写出程序，并通过计算机自动执行。

第十一节　复兴科技的核心竞争力

复兴科技早期在投资策略方面的技术，很可能运用的就是上述突破性的策略。通过HMM把股票收益分拆成三个分布以后，作者发现美国的数据和中国的数据一样，持续期可能只有1.5～1.6期。这样，就需要预先把单埋好，突破就买。由于不可能来回突破太多次，所以，每次买卖的量是有容量限制的。

复兴科技真正的核心竞争力在哪里呢？它的主要竞争力来源于对数学模型本身的深刻理解。比如，世界上真正理解HMM的人有多少？理解后，能够把它用算法实现出来的有多少人？HMM有很多假设，它的稳定性怎么样？这些估计值在多长时间里面是稳定的？如果这些估计值不稳定，策略是没有用的。复兴科技的顶级HMM专家一定有一套办法来判断所得到的参数值是否稳定，拆出来的分布是否稳定，他们在这些方面有自己的独到之处。

复兴科技的交易品种很多，同时对上万个合约、个股和指数进行交易。所以，这就是为什么复兴科技的计算机技术很厉害的原因，他们整个公司的运营全靠计算机，主要资产也都是计算机，没有其他的固定资产。同时，这也说明复兴科技公司的金融计算量是巨大的。

最后，我们应当认识到，对于任何复杂系统问题，都应分为两个步骤来解决：首先是找到解决问题的方法；其次，在此基础上，对解决办法进行简化和优化。任何方法过度复杂的话，一定会在实际应用中出现各种各样的问题。实际上，复兴科

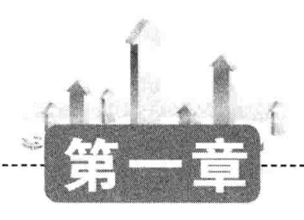
技采用的不应该是很复杂的办法，应该是一些相对简单的办法。当一个解决问题的办法出现了以后，紧跟着的问题是有没有更好、更简化的办法。

我们上面举的交易策略例子，只是复兴科技很多交易策略中的一个。复兴科技真正的核心在于基于数学模型及其计算结果来进行相应的交易策略设计。比如，上面的例子中，就是将 HMM 的计算结果转换成 Larry Williams 的突破策略。当股价突破的时候，通过 HMM 方法做到心中有数：突破后还剩几期，还有多长时间去加仓或者减仓。

接下来，我们首先介绍马尔可夫链的相关基础知识。

第一部分

基础知识



第一章 极大似然估计法简介

极大似然估计法（Maximum Likelihood Estimation Method）的概念主要是基于如下事实：不同的统计总体会产生出不同的样本，对于某一特定的样本，观察者很可能并不知道产生这一样本的确切的总体分布，但是该样本来自一些形式总体的可能性要比来自另一些的可能性大，即一些总体比另一些总体更容易产生出我们所观察到的样本。举例来说，假设我们抽取到了一个如图 1－1 所示的样本（$x_1, x_2, \cdots, x_8$），并且我们知道这一样本来自一个正态总体，同时假设我们也知道这个正态总体的方差，但却不知该分布的期望。假定这 8 个样本观察值不是来自 A 分布就是来自 B 分布，那么很明显，如果产生样本的真正分布是 B，那么我们观察到 $x_1, x_2, \cdots, x_8$ 这 8 个样本点的概率是非常小的。相反，如果真正的总体分布是 A，那么我们获得上述样本的可能性会显著增大。很显然，我们愿意接受 A 为真实的总体分布，因为 A 比 B 更可能产生出我们获得的样本观察值。在某种意义上，是样本“替”我们“选择”了总体分布 A。通常所说的“让数据说话”（let data talk）就是这个道理。

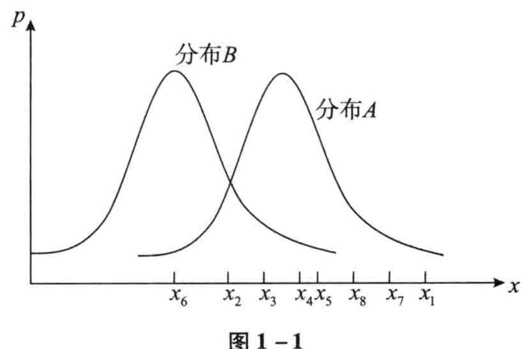


图 1－1

假定现在要根据从总体 ξ 中抽取到的样本 $(x_1,\cdots,x_n)$ ，对总体分布中的未知数 θ 进行估计。极大似然法选择使观察结果即样本 $(x_1,\cdots,x_n)$ 出现的概率最大的 $\hat{\theta}$ 作为 θ 的估计值。对于离散型随机变量，就是要选择使 $P(x_1)P(x_2)\cdots P(x_n)$ 最大的 $\hat{\theta}$ ；而对于连续型随机变量，就是要选择使 $\varphi(x_1)\varphi(x_2)\cdots\varphi(x_n)$ 最大的 $\hat{\theta}$ 。对于连续型随机变量，$\varphi(x_i)$ 表示随机变量在 x_i 附近取值的概率大小，因而相当于离散型随机变量中的 $P(x_i)$ 。

下面，我们用数学语言对上述思想进行说明。

设 ξ 为连续型随机变量，它的分布函数是 $F(x;\theta)$ ，分布密度是 $\varphi(x;\theta)$ ，其中 θ 是未知参数。由于抽取样本的独立性，则样本 $(x_1,\cdots,x_n)$ 的联合分布密度是：

$$L(x_1,\cdots,x_n;\theta) = \prod_{i=1}^{n}\varphi(x_i;\theta)$$

由于每个取定的样本值 $x_1,\cdots,x_n$ 是常数，所以 L 可看成参数 θ 的函数。我们把 L 称为样本的似然函数。若 ξ 为离散型随机变量，有概率函数 $P(\xi = x_i) = P(x_i;\theta)$ ，则似然函数 $L(x_1,\cdots,x_n;\theta) = \prod_{i=1}^{n}P(x_i;\theta)$ 。

定义 1.1　如果 $L(x_1,x_2,\cdots,x_n;\theta)$ 在 $\hat{\theta}$ 处取最大值，则称 $\hat{\theta}$ 是 θ 的极大似然估计。

为了求得 θ 的极大似然估计，我们必须使 L 达到最大值，并且把此时的 $\hat{\theta}$ 作为 θ 的估计量。由于 L 与 $\ln L$ 同时达到最大值，所以我们只需求 $\ln L$ 的最大值点即可，这样往往会给计算带来极大的方便。

由于 L（即 $\ln L$）是参数 θ 的函数，根据微积分中的拉格朗日定理，$\ln L$ 的最大值应在 $\ln L$ 对 θ 的一阶导数等于 0 时取到。因而，考虑方程 $\frac{d\ln L}{d\theta} = 0$ ，这个方程称为似然方程，容易看出，我们所要求的 $\hat{\theta}$ 就是这个似然方程中 θ 的解。

下面的两个例子可以帮助读者进一步了解极大似然法。

[例1－1] 已知随机变量ξ的分布为：

$$\xi \sim \varphi(x_i,\theta) = \begin{cases} \frac{1}{\theta}e^{-\frac{x}{\theta}} & x > 0(\theta > 0) \\ 0 & \text{其它} \end{cases}$$

$x_1,x_2,\cdots,x_n$是ξ的一组观察值，求θ的极大似然估计。

解 构造似然函数

$$L(x_1,\cdots,x_n;\theta) = \prod_{i=1}^{n} \frac{1}{\theta}e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^n} \cdot e^{-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$$

两边取对数，得：$\ln L = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n}x_i$

对θ求导，得：$\frac{d\ln L}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}x_i$

解似然方程：$\frac{d\ln L}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}x_i = 0$

可得：$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$，$\hat{\theta}$是$\theta$的极大似然估计。

[例1－2] 已知ξ服从正态分布$N(\mu,\sigma^2)$，$x_1,x_2,\cdots,x_n$为ξ的一组样本观察值，用极大似然法估计μ,σ^2的值。

解 $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L = \frac{n}{2}\ln\left(\frac{1}{2\pi}\right) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

由于有两个参数，这里应分别将L对μ和σ^2求偏导数。

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

这里要解如下的似然方程组

$$\begin{cases} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \\ -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=0 \end{cases}$$

解得：

$$\begin{cases} \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{cases}$$

通过以上例子，我们知道：

1. 当不只有一个总体分布参数需要估计时，应将 L 分别对各个不同的参数求偏导，然后解一个似然方程组。

2. 用极大似然法求出的总体方差的估计量 $\hat{\sigma}^2$ 是 $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$，在数理统计中，我们知道 $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 不是 σ^2 的无偏估计量，$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 才是 σ^2 的无偏估计量。事实上，用极大似然法对方差进行估计所得到的估计量往往都是有偏的，这一点在以后用极大似然法对线性回归模型进行估计时会再次说明。

第一节　线性模型的极大似然估计量

我们以简单的线性回归模型为基础，求解以下线性回归模型的极大似然估计量：

$$y_t=\beta_1+\beta_2x_t+u_t \qquad (1.1)$$

假设 $u_t \sim N(0,\sigma^2)$，$y_t \sim N(\beta_1+\beta_2x_t,\sigma^2)$，因此该正态分布的概率密度函数由下式给出：

$$f(y_t \mid \beta_1+\beta_2x_t,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(y_t-\beta_1-\beta_2x_t)^2}{\sigma^2}\right) \qquad (1.2)$$

由于 $y_t(t=1,2,\cdots,T)$ 是独立分布的，所以，所有 y_t 的联合概率密度函数可以

表示为单个概率密度函数的乘积：

$$
\begin{aligned}
& f(y_1, y_2, \cdots, y_T \mid x_1, x_2, \cdots, x_T, \beta_1, \beta_2, \sigma^2) \\
& = f(y_1 \mid \beta_1 + \beta_2 x_1, \sigma^2) f(y_2 \mid \beta_1 + \beta_2 x_2, \sigma^2) \cdots f(y_T \mid \beta_1 + \beta_2 x_T, \sigma^2) \\
& = \prod_{t=1}^{T} f(y_t \mid \beta_1 + \beta_2 x_t, \sigma^2) \qquad (1.3)
\end{aligned}
$$

该表达式中第一个等号左边被称为联合密度函数，右边被称作边际密度的乘积。基本的概率知识告诉我们，该结果成立的前提是 y 值具有独立性。对于三个独立的事件 A、B、C，其同时发生发概率是：A 发生的概率乘以 B 发生的概率再乘以 C 发生的概率。式 1.3 表示得到所有 y 值的实际概率，将每个式 1.2 中的 y_t 值代入式 1.3，并利用：

$$
Ae^{(x_1)} \times Ae^{(x_2)} \times \cdots \times Ae^{(x_T)} = A^T (e^{x_1} \times e^{x_2} \times \cdots \times e^{x_T}) = A^T e^{(x_1 + x_2 + \cdots + x_T)}
$$

得到如下结果：

$$
\begin{aligned}
& f(y_1, y_2, \cdots, y_T \mid x_1, x_2, \cdots, x_T, \beta_1, \beta_2, \sigma^2) \\
& = \frac{1}{\sigma^T (\sqrt{2\pi})^T} \exp\left(-\frac{1}{2} \frac{\sum (y_t - \beta_1 - \beta_2 x_t)^2}{\sigma^2}\right)
\end{aligned}
$$

这是所有在给定 x_t，β_1，β_2 和 σ^2 情况下 y_t 的联合密度，其中 $\sum$ 表示 $\sum_{t=1}^{T}$。然而，实际中发生的情况与上述过程相反，即给出 x_t 和 y_t，需要估计 β_1，β_2 和 σ^2。以上这种情况中的 $f(\cdot)$ 即为似然函数，记作 $LF(\beta_1, \beta_2, \sigma^2)$。

$$
LF(\beta_1, \beta_2, \sigma^2) = \frac{1}{\sigma^T (\sqrt{2\pi})^T} \exp\left(-\frac{1}{2} \frac{\sum (y_t - \beta_1 - \beta_2 x_t)^2}{\sigma^2}\right)
$$

极大似然估计的原理是要是选择参数 β_1，β_2，σ^2 的取值，使得极大似然函数 LF 的概率达到最大值。

对于单调函数 $f(x)$ 而言，$\max[f(x)] = \max[\ln f(x)]$。由于对数函数是单调函数，定义 $LLF = \ln LF$，所以 LF 和 LLF 在同一点达到最大，即在这两种情况下参数的最优值是一样的。这样，我们就可以对极大似然函数取对数，将其转化为 LLF 函数。由原 LF 取对数可以得到：

$$LLF = -T\ln\sigma - \frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(y_t - \beta_1 - \beta_2 x_t)^2}{\sigma^2}$$

为了使 σ^2 作为一个整体出现在方程中，我们将上式中的第一部分作了简单的转化，可得：

$$LLF = -\frac{T}{2}\ln\sigma^2 - \frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(y_t - \beta_1 - \beta_2 x_t)^2}{\sigma^2}$$

由于 $\frac{\partial}{\partial x}\ln(x) = \frac{1}{x}$，所以对上式中的 β_1，β_2，σ^2 分别求一阶导数得到：

$$\frac{\partial LLF}{\partial\beta_1} = \frac{1}{2}\sum\frac{2(y_t - \beta_1 - \beta_2 x_t)}{\sigma^2} \tag{1.4}$$

$$\frac{\partial LLF}{\partial\beta_2} = \frac{1}{2}\sum\frac{2(y_t - \beta_1 - \beta_2 x_t)x_t}{\sigma^2} \tag{1.5}$$

$$\frac{\partial LLF}{\partial\sigma^2} = -\frac{T}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum\frac{(y_t - \beta_1 - \beta_2 x_t)^2}{\sigma^4} \tag{1.6}$$

令式 1.4—1.6 的值为 0，并用 $\hat{\beta}_1$，$\hat{\beta}_2$，$\hat{\sigma}^2$ 来表示对相关参数的极大似然估计量，由式 1.4 得：

$$\sum(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t) = 0$$

$$\sum y_t - \sum\hat{\beta}_1 - \sum\hat{\beta}_2 x_t = 0$$

$$\sum y_t - T\hat{\beta}_1 - \hat{\beta}_2\sum x_t = 0$$

$$\frac{1}{T}\sum y_t - \hat{\beta}_1 - \hat{\beta}_2\frac{1}{T}\sum x_t = 0$$

由于 $\frac{1}{T}\sum y_t = \bar{y}$，$\frac{1}{T}\sum x_t = \bar{x}$，所以带入方程后，可得 $\hat{\beta}_1$ 的估计量为：

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x} \tag{1.7}$$

由式 1.5 得：

$$\sum(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t)x_t = 0$$

$$\sum y_t x_t - \hat{\beta}_1\sum x_t - \hat{\beta}_2\sum x_t^2 = 0$$

$$\hat{\beta}_2 \sum x_t^{\ 2} = \sum y_t x_t - (\bar{y} - \hat{\beta}_2 \bar{x}) \sum x_t$$

$$\hat{\beta}_2 \sum x_t^{\ 2} = \sum y_t x_t - T\bar{x}\bar{y} + \hat{\beta}_2 T\bar{x}^2$$

$$\hat{\beta}_2 (\sum x_t^{\ 2} - T\bar{x}^2) = \sum y_t x_t - T\bar{x}\bar{y}$$

$$\hat{\beta}_2 = \frac{\sum y_t x_t - T\bar{x}\bar{y}}{\sum x_t^{\ 2} - T\bar{x}^2} \tag{1.8}$$

由式 1.6 得：

$$\frac{T}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^4} \sum (y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t)^2$$

$$\hat{\sigma}^2 = \frac{1}{T} \sum (y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t)^2$$

注意在上式等式右边表示的是残差（即用实际值减去拟合值得到），因此：

$$\hat{\sigma}^2 = \frac{1}{T} \sum \hat{u}_t^2 \tag{1.9}$$

这些公式与 OLS 估计相比如何呢？由于式 1.7 和式 1.8 与 OLS 估计是完全相同的，所以极大似然估计和 OLS 估计将得到相同的截距和斜率系数。然而，式 1.9 中对 $\hat{\sigma}^2$ 的估计与 OLS 估计是不同的。

第二节　极大似然估计法的几个重点问题

通过上面的介绍，我们知道极大似然估计是一种统计方法，用它和样本数据可以得到相关密度函数的参数。比起只使用二阶距的最小二乘法而言，极大似然估计法整合了模型中的所有信息。由于后面很多章节的参数估计在很大程度上依赖于极大似然估计法，所以，我们下面主要讨论极大似然估计法在实际应用中所遇到重点问题。

一、极大似然估计和协方差矩阵

在计量经济学中，我们通常会已知向量 $y_{1:T} = (y_1, y_2, \cdots, y_t)$，也就是样本数据，但不知道相关统计模型中的参数向量 θ。在这种情况下，极大似然函数可以写为：

$$L = (\theta \mid y_{1:T})$$

从上式可以看出，不同的 θ 值会导致不同的似然估计函数值。我们在前面说过，参数的极大似然估计量，可以通过对似然函数取自然对数并求其最大值得到：

$$\hat{\theta}_{ML} = \mathrm{argmax} \ln L(\theta \mid y_{1:T})$$

其中，ln 表示对似然函数取自然对数值。

解上述对数似然函数最大化问题，不仅可以使我们得到参数的极大似然估计 $\hat{\theta}_{ML}$，还可以直接估计该函数的渐进协方差矩阵 $\mathrm{Cov}(\hat{\theta}_{ML})$。对该似然函数求二阶导数后的期望值，即为信息矩阵 $I(\theta)$：

$$I(\theta) = -E\left[\frac{\partial^2 \ln L(\theta \mid y_{1:T})}{\partial\theta\partial\theta'}\right]$$

信息矩阵 $I(\theta)$ 汇总了样本中的主要信息量。此信息矩阵的逆为我们提供了协方差矩阵的无偏估计 $\hat{\theta}$ 的下限值，该式子被称作 Cramer - Rao 不等式：

$$\mathrm{Cov}(\hat{\theta}) - I(\theta)^{-1} \geq 0$$

另外，也可以证明，该极大似然估计量 $\hat{\theta}_{ML}$ 的渐进分布为正态分布：

$$\sqrt{T}(\hat{\theta}_{ML} - \theta) \rightarrow N(0, (\bar{\mathrm{H}})^{-1})$$

$$-\frac{1}{T}\frac{\partial^2 \ln L(\theta \mid y_{1:T})}{\partial\theta\partial\theta'} \rightarrow \bar{\mathrm{H}} = \lim \frac{1}{T} I(\theta)$$

上面的公式为我们提供了如何通过对数似然函数的二阶导数，来对 $\hat{\theta}_{ML}$ 协方差矩阵进行估计的方法：

$$\mathrm{Cov}(\hat{\theta}_{ML}) = \left[\frac{\partial^2 \ln L(\theta \mid y_{1:T})}{\partial\theta\partial\theta'} \Big|_{\theta=\hat{\theta}_{ML}}\right]^{-1}$$

二、参数约束

极大似然估计值 $\hat{\theta}_{ML}$ 可以通过使对数似然方程等于0计算得到：

$$\frac{\partial \ln L(\theta|y_{1:T})}{\partial\theta}=0$$

大多数情况下，没有办法得到上面方程的解析解。因此，必须使用非线性的数值算法求最大化问题：给定最初的参数估计 θ^{j-1}，新的估计值 θ^{j} 可以通过关于 θ^{j-1} 的对数似然方程一阶偏导求得；新的极大似然估计值会大于原来的估计值；不断重复这个过程直到参数估计值收敛。这样，我们便可以得到上面这个方程的最优解。需要指出的是，在某些情况下，最优解不是唯一的。本书后面的一些章节会详细介绍该数值算法。

在无约束参数的极大似然函数求极值的时候，计算机会从负无穷到正无穷的参数空间中搜索。但是实际情况是，一些参数可能需要被约束在一定的区间范围之内。例如，如果 θ 中的一个参数表示概率 p，那么它的约束条件是 $0<p<1$。一般而言，这样的约束可以表示为一个无约束变量 ψ 的转化形式：

$$\theta=g(\psi)$$

其中，$g(\cdot)$ 是连续函数。

对数似然方程将写成：

$$\ln L(\theta|y_{1:T})=\ln L(g(\psi)|y_{1:T})\ln L(\psi|y_{1:T})$$

这样，就可以用未被约束的数值最优条件求解。

下面是我们经常遇到的三种情况：

1. 如果 θ_j 是 θ 的第j个元素，代表一个方差，那么 $\theta_j>0$。我们利用以下转化形式：

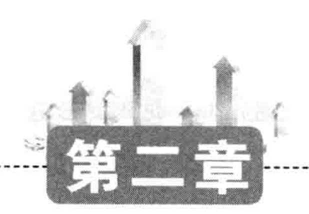
$$\theta_j=\psi_j^2 \text{ 或者 } \theta_j=\exp(\psi_i)$$

2. 如果 θ_j 表示一个概率，那么约束条件为：$0<\theta_j<1$，转化形式则是：

$$\theta_j = \frac{1}{1 + \exp(\psi_j^{-1})};$$

3. 如果 θ_j 表示一个 $AR(1)$ 模型中的自回归参数，那么它的约束条件是：$-1 < \psi_j < 1$，相应的转化式为：

$$\theta_j = \frac{\psi_j}{1 + |\psi_j|}$$



第二章

贝叶斯分析

第一节 统计学历史发展简介

统计学自从诞生以后大致沿着两条主线展开。一条主线是概率论，以 Blaise Pascal（1623—1662）和 Pierre Fermat（1601—1665）为先驱，最初用来解决计算赌博中的期望值和不确定性问题。此后概率论从数学的角度得到完善，Christian Huygens（1629—1695）、James Bernoulli（1654—1705）、Pierre - Simon Laplace（1749—1827）等都做出了巨大的贡献。另外，概率论从逻辑角度也得到了发展，Thomas Bayes（1701—1761）、George Boole（1815—1864）和 John Venn（1834—1923）对这一领域的研究贡献很大。经过不断发展，概率论成为科学史上一个重要的里程碑。Ronald Aylmer Fisher（1890—1962）认为，古希腊和伊斯兰数学家在概率方面所知甚少，甚至也有人认为人类的大脑不能够解决概率方面的问题。概率论脱胎于数学理论，并首次能够将不确定事件给予严密的表述。同时，概率论作为一种演绎研究方法，擅长以公理为基础，以假设为条件来得到事件发生概率方面的判断和推论，从而排除了具体观测值的影响，所以只能成为数学研究的一个分支。在这之后，概率论孕育了统计理论，以 1763 年 Thomas Bayes 论文发表为标志，现代统计学正式诞生，并被 Pierre - Simon Laplace 发展完善。所以，第一条主线是指在概率论的基础上产生的统计学。

另一条主线是几乎与之平行发展的误差理论。与概率论的主线不同，该理论的

重点不在于计算概率和不确定性的大小，而在于总结天文学和测量学中的观测数据。Carl Friedrich Gauss（1777—1855）是这个领域的主要贡献者，尤其是因为他提出了简单实用的估计方法，即最小二乘原理。误差理论发展的重要条件就是要有丰富的数据。Ronald Fisher 就曾指出，从现代统计诞生到 19 世纪末，Francis Galton（1822—1911）起到关键作用。作为一个对数据非常痴迷的数据搜集者和研究者，Francis Galton 坚持认为定量和统计方法在解决不确定性问题中是非常有力的工具。

在这之后，统计学的发展依然依赖于数据丰富的研究环境。这也就是为什么统计学被广泛应用于农业科学、医学和生物科学等领域，在该方向的研究中 Ronald Fisher（1890—1962）做出了很大的贡献。后来，统计学又被应用于诸多领域，包括质量控制、军事、工程、心理学、商业经济、公共政策和经济政策等等。而最令人着迷同时也具有广大发展空间的应用领域莫过于对投资的分析。

一、贝叶斯学派和频率学派

根据对不确定问题的不同回答，尤其是考虑概率方式的不同，统计学研究分为贝叶斯学派和频率学派两种。18、19 世纪的早期研究者认为，概率包括主观和客观两个层面的含义，前者指的是对一个事件发生的相信程度，后者指的是事件在长期、反复发生中所体现出的频率。到了 20 世纪，这个观点产生了巨大的分歧并出现了两个截然不同的学派，即频率学派和贝叶斯学派。频率学派认为概率只是长期重复实验所体现出来的频率，而贝叶斯学派认为概率可以包含对不确定性的主观见解。两者根本分歧在于，贝叶斯学派认为特定的情况及数据能够用来研究事件发生的概率，而频率学派则强调概率是长期实验频率的客观性结果。

举例来说，如果投掷一枚硬币，我们会怀有不确定性的主观认识，认为头像朝上的概率是 0.5。现在需要考虑的问题是：对于下一次投掷，我们可以主观上认为不确定性是 0.5？或者 0.5 只是代表了一个长期均值？贝叶斯学派认为两个解释都是有效的，而真正的频率学派只认同后者。由于两个学派对概率理论的不同认识而产生了不同的研究方法，这也反映了二者存在实质上的重要区别。这些争论表明统

计学的根基还不牢固，但是这并没有阻碍统计的应用，反而有利于统计学的辩证发展。

二、贝叶斯学派和逆概率

Thomas Bayes 在 1763 年发表的《机遇理论中一个问题的解》中，利用已观测数据，首次提出了定量归纳推理的现代方法。他用一个逆概率公式来估计一个二项式概率，也就是后来被广泛使用的标准方法——贝叶斯定理。

贝叶斯定理最简单的形式是：对于两个事件 A 和 B，有

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}$$

假设未知的二项式概率是 θ，在 n 次独立实验中观察到成功的次数是 x。那么，贝叶斯解法可以写为

$$f(\theta \mid x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x \mid \theta)f(\theta)}{\int f(x \mid \theta)f(\theta)\,d\theta}$$

其中，$f(\theta \mid x)$ 是给定条件 x 下 θ 发生的概率，$f(\theta)$ 是 θ 的先验概率，$f(x)$ 是 x 的边际概率。

第二节　贝叶斯分析简介

贝叶斯方法提供了一个在不确定环境下进行统计推断的有效工具。贝叶斯方法引入了一种对概率的全新阐释，即认为“概率”是对不确定性的一个条件测度。下面主要介绍贝叶斯分析的一些基本概念，包括贝叶斯理论、先验概率、后验概率等等。

一、贝叶斯定理

概率是描述事件发生可能性的有效方法。迄今为止，概率的概念已经被广泛应用于社会科学、自然科学、医学科学等诸多研究领域，在决策、预测以及随机结构挖掘等问题上发挥重要作用。在概率论中，我们对一个事件的概率可以做出两种解释：一种解释认为，概率是在一大批结果中事件发生频率的客观概率或频率概率；另一种解释认为，概率代表在给定信息和先验认识条件下，对特定事件发生的不确定性的一种条件测度，即贝叶斯概率。

我们用两个经典的例子进一步阐明概率的概念。首先，考虑抛掷一枚均匀硬币的问题。由于硬币是均匀的，硬币正面朝上和反面朝上的概率应当相等，均为0.5。这个概率是一个客观概率，它是根据一个事件发生的频率或者根据逻辑来定义的。其次，考虑另外一个问题：圆周率 $\pi=3.1415926\cdots\cdots$，它的第12位数是9的概率为多少？因为圆周率是一个确定性的数，所以它的第12位数并没有不确定性。因此，不能用客观概率的概念来分析这个问题，但是可以考虑用贝叶斯分析方法。贝叶斯分析中的概率依赖于先验知识，或者说以先验知识为条件。如果我们已经知道任意一个数是圆周率的第12位数的可能性都为1/10，那么，“圆周率的第12位数是9”这一判断为真的概率就只有1/10。这个概率就是一种贝叶斯概率，它表示对一个确定事件发生的相信程度的经验值，因此，同一问题中的贝叶斯概率可能是取不同的值。

贝叶斯理论是贝叶斯分析的基本工具，下面我们回顾一下贝叶斯理论：

假设 A 和 B 是两个确定事件，定义 $P(A|B)$ 为已知事件 B 已经发生的情况下事件 A 发生的概率。如果事件 B 发生的概率大于零，即 $P(B)>0$，那么在事件 B 发生的条件下事件 A 发生的条件概率为：

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \tag{2.1}$$

其中，$P(A\cap B)$ 为事件 A 与事件 B 同时发生的概率。给定事件 B，事件 A 的条

件概率随着概率 P（B）的变化而改变，这是因为我们考虑的是给定事件 B 已发生的情形。将上面关于条件概率的表达式变形，我们得到：

$$P(A \cap B) = P(A \mid B)P(B) \tag{2.2}$$

这个公式被称为**概率乘法规则**。

贝叶斯理论是以全概率法则为依据建立的。设 $A_1, A_2, \cdots, A_m$ 是不相交事件，即 $P(A_i \cap A_j) = 0, i \neq j$；并且 $P(A_1 \cup A_2 \cup \cdots \cup A_m) = 1$，设必然事件 $\Omega = A_1 \cup A_2 \cup \cdots \cup A_m$，$P(\Omega) = 1$。于是我们有：

$$P(B) = P(B \mid \Omega) = \sum_{j=1}^{m} P(B \mid A_j)P(A_j) \tag{2.3}$$

这样，必然事件 Ω 被分为 m 个不相交的子事件，事件 B 的条件概率即为给定每一个子事件 $A_j(j = 1,2,\cdots,m)$ 下事件 B 的条件概率之和。于是，给定事件 $B(P(B) > 0)$，事件 A_k 的条件概率可以写为：

$$P(A_k \mid B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k \cap B)}{\sum_{j=1}^{m} P(B \mid A_j)P(A_j)}$$

$$= \frac{P(B \mid A_k)P(A_k)}{\sum_{j=1}^{m} P(B \mid A_j)P(A_j)}, k = 1,2,\cdots,m \tag{2.4}$$

这个表达式就是**贝叶斯定理**。下面将介绍如何利用贝叶斯定理来进行基本的贝叶斯分析。

二、贝叶斯分析导论

我们在前面已经讨论过，贝叶斯方法将概率视为在给定信息和先验概率下，对特定事件发生的不确定性的一种条件测度。下面的例子将说明如何将贝叶斯定理应用到现实问题当中。

假设在总人口中有 5% 的人感染了某种病毒。我们从总人口中随机抽出一个人对其进行初步检验，结果为 Y，在该检验中被感染者测出阳性的概率为 97%，未被感染者测出阳性的概率为 30%。用事件 H_+ 表示被检测者携带该病毒，事件 H_- 表示

被检测者不携带该病毒。根据已上信息可得到：

$$P(H_+) = 0.05$$

$$P(H_-) = 0.95$$

同样，用事件 Y_+ 表示 X 检验结果为阳性，事件 Y_- 表示 X 检验结果为阴性。根据检验 Y 出现不同结果的概率可知：

$$P(Y_+ | H_+) = 0.97$$

$$P(Y_+ | H_-) = 0.30$$

这里，$P(Y_+ | H_+) = 0.97$ 是在已知信息 H_+ 的条件下事件 Y_+ 发生的概率，也可理解为在已知信息 H_+ 的情况下检验 Y_+ 的精确度。

下面，我们关注的是在已知检验结果为阳性后，被测试者实际携带该病毒的概率 $P(H_+ | Y_+)$。根据贝叶斯理论，可以计算出：

$$\begin{aligned} P(H_+ | Y_+) &= \frac{P(Y_+ | H_+)P(H_+)}{P(Y_+)} \\ &= \frac{P(Y_+ | H_+)P(H_+)}{P(Y_+ | H_+)P(H_+) + P(Y_+ | H_-)P(H_-)} \\ &= \frac{0.97 \times 0.05}{0.97 \times 0.05 + 0.30 \times 0.95} \\ &\approx 0.15 \end{aligned}$$

从以上结果可以看出，给定 Y 检验为阳性（Y_+）这一信息后，该人携带该病毒的概率由 5% 提高到了 15%。

现在，我们让该被测试者进行一个更加精确的检验 Q，这种病毒检验 Q 的精确度为：

$$P(Q_+ | H_+) = 0.98$$

$$P(Q_+ | H_-) = 0.10$$

在该被测试者进行检验 Y 之前，我们可以预测检验 Y 结果为阳性的概率为：

$$\begin{aligned} P(Q_+ | Y_+) &= P(Q_+, H_+ | Y_+) + P(Q_+, H_- | Y_+) \\ &= P(Q_+ | H_+) \times P(H_+ | Y_+) + P(Q_+ | H_-) \times P(H_- | Y_+) \end{aligned}$$

$$= 0.98 \times 0.15 + 0.10 \times 0.95$$

$$\approx 0.24$$

并且 $P(Q_-| Y_+) = 1 - P(Q_+| Y_+) \approx 0.76$。

下面，我们关注的是给定 Y 检验为阳性且 Q 检验为阴性下，被测试者携带该病毒的概率 $P(H_+| Y_+, Q_-)$。

由于　$P(A \cap B | C) = P(A | B \cap C)P(B | C) = P(B | A \cap C)P(A | C)$

令 $A = H_+, B = Q_-, C = Y_+$，可得：

$$\begin{aligned} & P(H_+, Q_-| Y_+) \\ &= P(H_+| Y_+, Q_-)P(Q_-| Y_+) \\ &= P(Q_-| H_+, Y_+)P(H_+| Y_+) \\ &= P(Q_-| H_+)P(H_+| Y_+) \end{aligned}$$

其中最后一个等式成立是因为 $P(Q_-| H_+, Y_+) = P(Q_-| H_+)$。从而可得：

$$P(H_+| Y_+, Q_-) = \frac{P(Q_-| H_+)P(H_+| Y_+)}{P(Q_-| Y_+)}$$

$$= \frac{P(Q_-| H_+)P(H_+| Y_+)}{P(Q_-| H_+)P(H_+| Y_+) + P(Q_-| H_-)P(H_-| Y_+)}$$

$$= \frac{0.02 \times 0.15}{0.02 \times 0.15 + (1 - 0.10) \times (1 - 0.15)}$$

$$\approx 0.0038$$

因此，检验 Q 的结果将被测试者携带病毒的概率从15%降低到了0.38%。

综上所述，一个人携带该病毒的概率随着给定信息的不同而发生以下变化：

$$P(H_+| \text{信息}) = \begin{cases} 5\% & \text{进行 } Y \text{ 检验和 } Q \text{ 检验之前} \\ 15\% & Y \text{ 检验阳性}, Q \text{ 检验之前} \\ 0.38\% & Y \text{ 检验阳性}, Q \text{ 检验阴性} \end{cases}$$

在进行观察之前，我们就对被测试者携带该病毒的概率有一个先验概率。在得到 X 检验的观测值后，通过计算后验概率对被测试者携带该病毒的概率进行修正。此外，我们还可以预测 Q 检验呈阳性的可能性大小。最后，我们将检验 Q 的结果也

纳入考虑，再次修正这一概率。也就是说，我们可以适当将所有可以得到的信息纳入考虑，从而修正被测试者携带该病毒的概率。

我们也可以从下面的角度来求 $P(H_+ | Y_+, Q_-)$：给定 Y 检验阳性、Q 检验阴性，我们可以将检验（Y, Q）联合起来看，计算出的 $P(H_+ | Y_+, Q_-)$。由于检验 Y 和检验 Q 是相互独立的，我们有：

$$P(Y_+, Q_- | H_+) = P(Y_+ | H_+)P(Q_- | H_+) = 0.0194$$

$$P(Y_+, Q_- | H_-) = P(Y_+ | H_-)P(Q_- | H_-) = 0.27$$

根据贝叶斯理论得到：

$$\begin{aligned}P(H_+ | Y_+, Q_-) &= \frac{P(Y_+, Q_- | H_+)P(H_+)}{P(Y_+, Q_- | H_+)P(H_+) + P(Y_+, Q_- | H_-)P(H_-)} \\ &= \frac{0.0194 \times 0.05}{0.0194 \times 0.05 + 0.27 \times 0.95} \\ &\approx 0.0038\end{aligned}$$

这表明在给定信息下，两种方法的结果都是基本相同的。下面，我们分别运用这两种方法计算 $P(H_+ | Y_+, Q_+)$：

方法一：

$$\begin{aligned}P(H_+ | Y_+, Q_+) &= \frac{P(Q_+ | H_+)P(H_+ | Y_+)}{P(Q_+ | Y_+)} \\ &= \frac{P(Q_+ | H_+)P(H_+ | Y_+)}{P(Q_+ | H_+)P(H_+ | Y_+) + P(Q_+ | H_-)P(H_- | Y_+)} \\ &= \frac{0.98 \times 0.15}{0.98 \times 0.15 + (0.10) \times (1 - 0.15)} \\ &\approx 0.6248\end{aligned}$$

方法二：

$$P(Y_+, Q_+ | H_+) = P(Y_+ | H_+)P(Q_+ | H_+) = 0.9702$$

$$P(Y_+, Q_+ | H_-) = P(Y_+ | H_-)P(Q_+ | H_-) = 0.03$$

根据贝叶斯定理得到：

$$P(H_+ | Y_+, Q_+) = \frac{P(Y_+, Q_+ | H_+)P(H_+)}{P(Y_+, Q_+ | H_+)P(H_+) + P(Y_+, Q_+ | H_-)P(H_-)}$$

$$= \frac{0.9702 \times 0.05}{0.9702 \times 0.05 + 0.03 \times 0.95}$$

$$\approx 0.6299$$

以上分析表明，两种方法结果基本一致，但由于计算过程小数保留位数的原因造成些许误差。

当然，我们也可以进行两次 Y 实验。下面，假设我们没有 Q 检验，仅对同一个被测试者进行两次 Y 检验，结果分别为阳性、阴性（记为 Y_+^1, Y_-^2），假设两次检验是独立的。于是：

$$P(Y_+^1, Y_-^2 | H_+) = P(Y_+^1 | H_+)P(Y_-^2 | H_+) = 0.97 \times 0.03 = 0.0291$$

$$P(Y_+^1, Y_-^2 | H_-) = P(Y_+^1 | H_-)P(Y_-^2 | H_-) = 0.3 \times 0.7 = 0.21$$

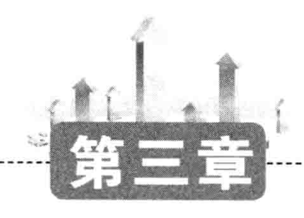
根据贝叶斯理论得到：

$$P(H_+ | Y_+^1, Y_-^2) = \frac{P(Y_+^1, Y_-^2 | H_+)P(H_+)}{P(X_+^1, X_-^2 | H_+)P(H_+) + P(Y_+^1, Y_-^2 | H_-)P(H_-)}$$

$$= \frac{0.0291 \times 0.05}{0.0291 \times 0.05 + 0.21 \times 0.95}$$

$$\approx 0.007240$$

经过两次 Y 检验，被测试者携带该病毒的概率由一次 Y 试验后的 15% 减小到了约 0.7240%，虽然没有进行 Y 检验后再进行 Q 检验精确，但仍然有助于对被测试者是否携带该病毒进行判断。但是，现实中两次 Y 检验的结果很可能不是相互独立的，因而实际上给定 Y_+^1, Y_-^2 信息后，被测试者携带病毒的条件概率应该比 0.7240% 要大，但这仍然为提高检验精确度提供了一种思路。这也就是医生在很多情况下要求病人复查的原因。



第三章

马尔科夫链

根据实际研究的需要，我们经常要将一个系统分为有限个状态，甚至分为可数无限状态，然后根据所处状态的不同来对系统进行研究。令 $\mathfrak{M}$ 表示这样状态的集合，具体地，$\mathfrak{M}$ 可以取整数集的子集，则 $\mathfrak{M}$ 就是这个系统的状态矢量空间。假设在离散的时间点上对系统进行观测，并令 X_t 表示系统在 t 时刻的状态。由于我们的研究对象是非确定性的系统，所以对不同的时间点来说，通常假设 $X_t(t \geq 0)$ 是定义在同一概率空间的随机变量。尽管有了以上假设，但问题依然比较复杂，仍需要在其他方面做出假设。

比如，可以假设 X_t 是相互独立的随机变量。对于一个不断重复的系统来说，人们常常假设该系统未来所处状态与现在和过去都不相关，这样的独立性假设会使问题的处理得到相当程度的简化。但在现实中独立性未必得到满足，即便系统在过去和现在所处的状态不会直接确定未来状态，但也会对未来所处的状态产生一定影响。马尔科夫对独立性这个较强的假定做出了适当的放松，相对于独立性条件来说更加贴近现实。

马尔科夫性是指在当期取值给定时，滞后期取值将不会对未来值产生影响。具有这种特性的系统被称为**马尔科夫链**。马尔科夫性如式 3.1 所示：

$$P(X_{t+1} = x_{t+1} \mid X_0 = x_0, \cdots, X_t = x_t) = P(X_{t+1} = x_{t+1} \mid X_t = x_t) \qquad (3.1)$$

$P(X_{t+1} = y \mid X_t = x)$ 是马尔科夫链的转移概率。接下来，将介绍具有固定转移概率的马尔科夫链，也就是 $P(X_{t+1} = y \mid X_t = x)$ 与 t 无关的情况。下文中如果说到 $X_t(t \geq 0)$ 是一个马尔科夫链，那么则意味着这些随机变量满足马尔科夫性并且有固

定的转移概率。研究这种特殊的马尔科夫链是因为它有很强的理论基础，而且这些理论容易让初学者接受。另外，现实中有大量的系统可被归为马尔科夫链，它在实际操作中非常实用。

首先，考虑只有两种状态的马尔科夫链。

第一节　有两种状态的马尔科夫链

举一个有关汽车的例子。汽车每天都可能出现故障，假设在第 t 天出现故障并在第（$t+1$）天修好的概率为 p；在第 t 天运转正常并且在第（$t+1$）天出现故障的概率为 q。

π_0（0）是汽车在第 0 天出现故障的概率。状态 0 代表出现故障，状态 1 代表运转正常。随机变量 X_t 代表汽车在第 t 天的运转状态。则有：

$$P(X_{t+1}=1 \mid X_t=0)=p$$

$$P(X_{t+1}=0 \mid X_t=1)=q$$

$$P(X_0=0)=\pi_0(0)$$

由于系统只有 0 和 1 两种状态，根据上式可以得到：

$$P(X_{t+1}=0 \mid X_t=0)=1-p$$

$$P(X_{t+1}=1 \mid X_t=1)=1-q$$

并且汽车的初始状态为 1 的概率 π_0（1）为：

$$\pi_0(1)=P(X_0=1)=1-\pi_0(0)$$

下面，我们根据这些信息计算 $P(X_{t+1}=0)$ 和 $P(X_{t+1}=1)$。

$P(X_{t+1}=0)$

$=P(X_t=0, X_{t+1}=0)+P(X_t=1, X_{t+1}=0)$

$=P(X_t=0)\ P(X_{t+1}=0 \mid X_t=0)+P(X_t=1)\ P(X_{t+1}=0 \mid X_t=1)$

$= (1-p)\,P(X_t=0) + q\,P(X_t=1)$

$= (1-p)\,P(X_t=0) + q\,(1-P(X_t=0))$

$= (1-p-q)\,P(X_t=0) + q$

且 $P(X_0=0)=\pi_0(0)$，所以 $P(X_1=0)=(1-p-q)\pi_0(0)+q$。并且，$P(X_2=0)=(1-p-q)P(X_1=0)+q=(1-p-q)^2\pi_0(0)+q[1+(1-p-q)]$。

把上式重复 t 次，得到：

$$P(X_t=0)=(1-p-q)^t\pi_0(0)+q\sum_{j=0}^{t-1}(1-q-p)^j \tag{3.2}$$

当 $p=q=0$ 时，对于所有的 t，都有：

$$P(X_t=0)=\pi_0(0) \qquad P(X_t=1)=\pi_0(1)$$

当 $p+q>0$ 时，根据等比数列求和公式：

$$\sum_{j=0}^{t-1}(1-p-q)^j=\frac{1-(1-p-q)^t}{p+q}$$

代入式 3.2，可以得到：

$$P(X_t=0)=\frac{q}{p+q}+(1-p-q)^t(\pi_0(0)-\frac{q}{p+q}) \tag{3.3}$$

$$P(X_t=1)=\frac{p}{p+q}+(1-p-q)^t(\pi_0(1)-\frac{p}{p+q}) \tag{3.4}$$

假设 p 和 q 不同时为 0 或 1，则 $0<p+q<2$，即 $|1-p-q|<1$。在这种情况下，我们令式 3.3 和式 3.4 中的 $t\to\infty$，得：

$$\lim_{t\to\infty}P(X_t=0)=\frac{q}{p+q}$$

$$\lim_{t\to\infty}P(X_t=1)=\frac{p}{p+q}$$

我们还可以用其他方法求 $\frac{q}{p+q}$ 和 $\frac{p}{p+q}$。选定 $\pi_0(0)$ 和 $\pi_0(1)$ 作为研究对象并且使 $P(X_0=0)$ 和 $P(X_t=1)$ 独立于 t，那么根据式 3.3 和式 3.4，得到：

$$\pi_0(0)=\frac{q}{p+q},\ \pi_0(1)=\frac{p}{p+q}$$

因此如果系统 $X_t(t \geq 0)$ 是从上述分布开始的，即：

$$P(X_0 = 0) = \frac{q}{p+q}, P(X_0 = 1) = \frac{p}{p+q}$$

那么对于系统中所有的 t ，都有：

$$P(X_t = 0) = \frac{q}{p+q}, P(X_t = 1) = \frac{p}{p+q}$$

上面的这个例子并没有清楚地说明该系统状态 $X_t(t \geq 0)$ 是否一定服从具有马尔科夫性。但是，当我们假设该系统状态服从马尔科夫性后，就可以计算 X_1，X_2，…，X_t 的联合分布。

令 $t = 2$, X_0, X_1 和 X_2 为0或1。那么：

$P(X_0 = x_0, X_1 = x_1, X_2 = x_2)$

$= P(X_0 = x_0, X_1 = x_1)P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$

$= P(X_0 = x_0)\ P(X_1 = x_1 \mid X_0 = x_0)\ P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$

上式中除 $P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$ 之外，$P(X_0 = x_0)$ 和 $P(X_1 = x_1 \mid X_0 = x_0)$ 都能由 p，q 和 $\pi 0(0)$ 表示。如果系统满足马尔科夫性，则有：

$$P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1) = P(X_2 = x_2 \mid X_1 = x_1)$$

上式由 q，p 取值决定。在这种情况下，

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, X_2 = x_2) \\ &= P(X_0 = x_0)\ P(X_1 = x_1 \mid X_0 = x_0)\ P(X_2 = x_2 \mid X_1 = x_1) \end{aligned}$$

根据以上公式，可以计算出以下各种情况的概率：

$P(X_0 = 0, X_1 = 0, X_2 = 0)$

$= P(X_0 = 0)\ P(X_1 = 0 \mid X_0 = 0)\ P(X_2 = 0 \mid X_1 = 0)$

$= \pi_0(0)\ (1 - p)^2$

$P(X_0 = 0, X_1 = 0, X_2 = 1)$

$= P(X_0 = 0)\ P(X_1 = 0 \mid X_0 = 0)\ P(X_2 = 1 \mid X_1 = 0)$

$= \pi_0(0)(1 - p)p$

$$P(X_0 = 0, X_1 = 1, X_2 = 0)$$
$$= P(X_0 = 0)\ P(X_1 = 1 \mid X_0 = 0)\ P(X_2 = 0 \mid X_1 = 1)$$
$$= \pi_0(0)pq$$

$$P(X_0 = 0, X_1 = 1, X_2 = 1)$$
$$= P(X_0 = 0)\ P(X_1 = 1 \mid X_0 = 0)\ P(X_2 = 1 \mid X_1 = 1)$$
$$= \pi_0(0)p(1 - q)$$

$$P(X_0 = 1, X_1 = 0, X_2 = 0)$$
$$= P(X_0 = 1)\ P(X_1 = 0 \mid X_0 = 1)\ P(X_2 = 0 \mid X_1 = 0)$$
$$= (1 - \pi_0(0))\ q(1 - p)$$

$$P(X_0 = 1, X_1 = 0, X_2 = 1)$$
$$= P(X_0 = 1)\ P(X_1 = 0 \mid X_0 = 1)\ P(X_2 = 1 \mid X_1 = 0)$$
$$= (1 - \pi_0(0))pq$$

$$P(X_0 = 1, X_1 = 1, X_2 = 0)$$
$$= P(X_0 = 1)\ P(X_1 = 1 \mid X_0 = 1)\ P(X_2 = 0 \mid X_1 = 1)$$
$$= (1 - \pi_0(0))(1 - q)q$$

$$P(X_0 = 1, X_1 = 1, X_2 = 1)$$
$$= P(X_0 = 1)\ P(X_1 = 1 \mid X_0 = 1)\ P(X_2 = 1 \mid X_1 = 1)$$
$$= (1 - \pi_0(0))\ (1 - q)^2$$

结果汇总如表 3－1。

表 3－1

X_0	X_1	X_2	$P(X_0 = x_0, X_1 = x_1, X_2 = x_2)$
0	0	0	$\pi_0(0)(1-p)^2$
0	0	1	$\pi_0(0)(1-p)p$
0	1	0	$\pi_0(0)pq$
0	1	1	$\pi_0(0)p(1-q)$
1	0	0	$(1-\pi_0(0))q(1-p)$

续表

X_0	X_1	X_2	$P(X_0 = x_0, X_1 = x_1, X_2 = x_2)$
1	0	1	$(1-\pi_0(0))qp$
1	1	0	$(1-\pi_0(0))(1-q)q$
1	1	1	$(1-\pi_0(0))(1-q)^2$

第二节 转移函数和初始分布

考虑放松系统只有两个状态的约束条件。假设 $X_t(t \geq 0)$ 是状态空间为 $\mathfrak{M}$ 的马尔可夫链，$x \in \mathfrak{M}$，$y \in \mathfrak{M}$，状态转移函数 $P(x,y)$ 可以定义为：

$$P(x,y) = P(X_1 = y \mid X_0 = x) \quad x,y \in \mathfrak{M} \tag{3.5}$$

且有：

$$P(x,y) \geq 0, \ x,y \in \mathfrak{M} \tag{3.6}$$

$$\sum_y P(x,y) = 1, \ x \in \mathfrak{M} \tag{3.7}$$

由于马尔可夫链中的状态转移概率是稳定的，我们可以得到：

$$P(X_{t+1} = y \mid X_t = x) = P(x,y) \quad t \geq 1$$

由马尔可夫性可以得出：

$$P(X_{t+1} = y \mid X_0 = x_0, \cdots, X_{t-1} = x_{t-1}, X_t = x)$$
$$= P(X_{t+1} = y \mid X_t = x)$$
$$= P(x,y)$$

也就是说，如果马尔可夫链在时点 t 的状态是 x，那么不管它过去如何达到 x，它在下一步中达到状态 y 的概率是 $P(x,y)$。因此，$P(x,y)$ 被称为马尔可夫链的**一阶转移概率**。

马尔可夫链的初始状态分布函数 $\pi_0(x)$（$x \in \mathfrak{M}$）被定义为：

$$\pi_0(x) = P(X_0 = x), \ x \in \mathfrak{M}$$

并且有：

$$\pi_0(x) \geq 0, x \in \mathfrak{M}$$

$$\sum_x \pi_0(x) = 1$$

在以上假设的基础上，可以得到 $X_0, \cdots, X_t$ 的联合分布关于转移函数和初始分布的函数。比如，

$P(X_0 = x_0, X_1 = x_1)$

$= P(X_0 = x_0)P(X_1 = x_1 \mid X_0 = x_0)$

$= \pi_0(x_0)P(x_0, x_1)$

同理，

$P(X_0 = x_0, X_1 = x_1, X_2 = x_2)$

$= P(X_0 = x_0, X_1 = x_1)P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$

$= \pi_0(x_0)P(x_0, x_1)P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$

由于 $X_t(t \geq 0)$ 满足马尔可夫性质并且有稳定的转移概率，所以：

$P(X_2 = x_2 \mid X_1 = x_1, X_0 = x_0)$

$= P(X_2 = x_2 \mid X_1 = x_1)$

$= P(x_1, x_2)$

因此，

$P(X_0 = x_0, X_1 = x_1, X_2 = x_2)$

$= P(X_0 = x_0, X_1 = x_1)P(X_2 = x_2 \mid X_0 = x_0, X_1 = x_1)$

$= \pi_0(x_0)P(x_0, x_1)P(x_1, x_2)$

归纳可得以下公式：

$P(X_0 = x_0, X_1 = x_1, \cdots, X_t = x_t)$

$= \pi_0(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_{t-1}, x_t)$

在以后的学习中，读者们将体会到以上公式的重要性。

第三节 马尔科夫链的一些性质

我们现在介绍马尔科夫链的一些性质，这是研究隐蔽马尔科夫模型（Hidden Markov Model，HMM）的基础。下面的讨论仅限于后面章节所需要的离散马尔科夫模型的研究。因此，尽管我们会提及不可约性（Irreducibility）和非周期性（Aperiodicity）等属性，但不会过多地讨论其技术层面的细节问题。

定义 3.1 **如果一个离散型随机变量序列** $\{C_t: t \in N\}$ **，** $C_t \in \{1,2,...,m\}$ **，对于所有的** $t \in N$ **都满足马尔科夫性质：** $P(C_{t+1} \mid C_t,\cdots,C_1) = P(C_{t+1} \mid C_t)$ **，则称此序列为离散时间的马尔科夫链。**

也就是说，以第 t 期及之前所有期的历史数据为条件来推断第 $t+1$ 期，等价于仅以最近一期的数据 C_t 为条件来推断第 $t+1$ 期。为了简化公式，我们定义 C_1^t 表示整个历史过程（$C_t,\cdots,C_1$），于是在此条件下马尔科夫过程可以表示为：$P(C_{t+1} \mid C_1^t) = P(C_{t+1} \mid C_t)$ 。

我们可以认为，马尔科夫性是对独立性假设的初步放松。为了数学上计算的方便，马尔科夫性假设随机变量 $\{C_t\}$ 仅仅依赖于前一期的数据，过去时期仅通过现在一期对未来一期产生影响。

在马尔科夫链中非常重要的概念和环节就是下面的条件概率，我们称之为**转移概率**，即在 s 时刻所处状态 i 转移到 $s+t$ 时刻所处状态 j 的概率：

$$P(C_{s+t} = j \mid C_s = i)$$

如果这些转移概率独立于时间 s，那么这个马尔科夫链被称为**齐次的**（Homogeneous），否则为**非齐次的**（Nonhomogeneous）。在不做特别说明的情况下，本书中所讨论的马尔科夫链都是齐次的，从而在此条件下转移概率可以表示为：

$$\gamma_{ij}(t) = P(C_{s+t} = j \mid C_s = i)$$

值得注意的是，上式中 $\gamma_{ij}(t)$ 与 s 无关，而由 $\gamma_{ij}(t)$ 所组成的转移矩阵 $\Gamma(t)$ 也与 s 无关，该矩阵中 (i,j) 位置的元素代表 $\gamma_{ij}(t)$。

有限状态空间的齐次马尔科夫链都具有一个十分重要的性质，即满足 Chapman-Kolmogorov 方程：

$$\Gamma(t+u)=\Gamma(t)\Gamma(u)$$

Chapman – Kolmogorov 方程的具体证明过程如下：

$$\Gamma(t)=\begin{pmatrix}\gamma_{11}(t) & \cdots & \gamma_{1m}(t)\\ \vdots & \ddots & \vdots\\ \gamma_{m1}(t) & \cdots & \gamma_{mm}(t)\end{pmatrix}$$

$$\Gamma(u)=\begin{pmatrix}\gamma_{11}(u) & \cdots & \gamma_{1m}(u)\\ \vdots & \ddots & \vdots\\ \gamma_{m1}(u) & \cdots & \gamma_{mm}(u)\end{pmatrix}$$

这两个矩阵相乘得到的新的矩阵的第 i 行第 j 列的元素为：

$$\begin{aligned}\sum_{k=1}^{m}\gamma_{ik}(t)\gamma_{kj}(u) &= \sum_{k=1}^{m}P(C_{s+t}=k\mid C_s=i)P(C_{s+u}=j\mid C_s=k)\\ &= \sum_{k=1}^{m}P(C_{s+t}=k\mid C_s=i)P(C_{(s+t)+u}=j\mid C_{s+t}=k)\\ &= \sum_{k=1}^{m}P(C_{s+(t+u)}=j\mid C_s=i)\end{aligned}$$

而 $\Gamma(t+u)$ 的第 i 行第 j 列的元素也为：$\sum_{k=1}^{m}P(C_{s+(t+u)}=j\mid C_s=i)$，因此 $\Gamma(t+u)=\Gamma(t)\Gamma(u)$ 得证。

根据 Chapman – Kolmogorov 等式，对所有 $t\in N$，有 $\Gamma(t)=\Gamma(1)^t$；也就是 t 步转移的转移概率矩阵（Transition Probability Matrix，简称“t. p. m.”）是 $\Gamma(1)$ 也就是第一步转移的转移概率矩阵的 t 次方。矩阵 $\Gamma(1)$（下面将会被简写为 Γ）是一个由概率组成的方阵，它每行的元素之和为 1：

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix}$$

其中，m 表示马尔科夫链的状态数。行的元素和为 1，即列向量 1’ 是 Γ 的特征向量并对应于特征值 1（这是因为 $\Gamma \cdot 1 = 1$）。我们可以视 Γ 为第一步转移的转移概率矩阵。请大家注意，这里向量表示不同于一般习惯，形如 1 的向量表示的是行向量。

马尔科夫链在给定时间 t 处于给定状态的无条件概率 $P(C_t = j)$ 是研究的重点。我们用下面行向量来表示上述概率：

$$u(t) = (P(C_t = 1), \cdots, P(C_t = m)), \ t \in N$$

$u(1)$ 表示马尔科夫链的初始分布（Initial Distribution）。为了通过 t 期的信息得到第$t+1$期的分布，我们把第 t 期概率分布右乘转移概率矩阵 Γ：

$$u(t+1) = u(t)\Gamma$$

这是由于：$P(C_{t+1} = i)$

$$= \sum_{j=1}^{m} P(C_t = j) P(C_{t+1} = i \mid C_t = j)$$

$$= \sum_{j=1}^{m} P(C_t = j) \gamma_{ji}$$

而 $u(t)$ 右乘 Γ 所得向量的第 i 个元素也为 $\sum_{j=1}^{m} P(C_t = j)\gamma_{ji}$，因此上式得证。

下面通过一个具体例子来说明。现在有一个描述股市是牛市或者熊市的时间序列数据，假设每天股市的状态仅与前一天的状态有关，转移概率矩阵如表 3－2。

表 3－2

		第 $t+1$ 天	
		熊市	牛市
第 t 天	熊市	0.8	0.2
	牛市	0.3	0.7

也就是说，如果今天是熊市，那么明天也将是熊市的概率是0.8；如果今天是牛市，那么明天是牛市的概率是0.7。于是该股市符合有两个状态的齐次马尔科夫链，其转移概率矩阵为：

$$\Gamma = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$$

用1代表熊市，2代表牛市，假设今天也就是第一期为牛市，这意味着今天股市的概率分布为：

$$u(1) = (P(C_1 = 1), P(C_1 = 2)) = (0 \quad 1)$$

明天和后天股市的概率分布可以通过对 $u(1)$ 连续右乘 Γ 得到：

$$u(2) = (P(C_2 = 1), P(C_2 = 2)) = u(1)\Gamma = (0.3 \quad 0.7)$$

$$u(3) = (P(C_3 = 1), P(C_3 = 2)) = u(2)\Gamma = (0.45 \quad 0.55)$$

……

如果对于一个转移概率矩阵为 Γ 的马尔科夫链，满足条件 $\delta\Gamma = \delta$ 并且 $\delta 1' = 1$ ，则称其有**稳态分布** δ , δ 为所有元素均为非负的行向量。其中第一个条件描述的是稳定性，而第二个条件则是要求 δ 确实为概率分布。马尔科夫链的转移矩阵举例如下：

$$\Gamma = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

其稳态分布为 $\delta = \frac{1}{32}(15 \quad 9 \quad 8)$ 。

可以根据定义中的两个条件来求出稳态分布，即要满足：

$$\begin{cases} (\delta_1 \quad \delta_2 \quad \delta_3)\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix} = (\delta_1 \quad \delta_2 \quad \delta_3) \\ \delta_1 + \delta_2 + \delta_3 = 1 \end{cases}$$

由于 $u(t+1) = u(t)\Gamma$ ，如果一个马尔科夫链从它的稳态分布开始，之后所有

时间点都会有同样的分布，我们把这样的过程叫做**稳态马尔科夫链**（Stationary Markov Chain）。需要说明的是，马尔科夫链具有齐次性并不是其成为稳态马尔科夫链的充分条件。因此我们将初始分布是稳态分布的齐次马尔科夫链（Homogeneous Markov chain）前面加一个形容词——“稳态”（Stationary）。不可约的（齐次、离散时间、有限状态的）马尔科夫链有唯一的、严格为正的稳态分布。

马尔科夫链的可逆性常常引起研究者的兴趣。如果一个随机过程的分布在逆转的时间下是不变的，那么称这个随机过程是可逆的。对于转移矩阵为Γ，稳态分布为 δ 的不可约的马尔科夫链，可逆性的一个充分必要条件是，对所有的状态 i 和 j：

$$\delta_i\gamma_{ij} = \delta_j\gamma_{ji}$$

两种状态的不可约的稳态马尔科夫链也满足上述条件，因此，这样的马尔科夫链是可逆的。

下面，我们对 HMM 的自相关函数（Auto - Correlation Function，简称“ACF”）进行比较。假设马尔科夫链是稳态的并且是不可化简的，那么它的 ACF 可以通过如下方法得到：

首先，定义 $\mathrm{v} = (1, 2, \cdots, m)$ 并且 $\mathrm{V} = diag(1, 2, \cdots, m)$，则对于所有非负整数 k，都有：

$$\mathrm{Cov}(C_t, C_{t+k}) = \delta \mathrm{V}\Gamma^k \mathrm{v}' - (\delta \mathrm{v}')^2$$

第二，如果Γ可对角化，并且它除了 1 之外的特征值表示为 $\omega_2, \omega_3, \cdots, \omega_m$，那么Γ可以被写成 $\Gamma = \mathrm{U}\Omega\mathrm{U}^{-1}$，其中是 $\Omega = diag(1, \omega_2, \omega_3, \cdots, \omega_m)$，并且 U 的各列向量是Γ的对应特征值的特征向量。于是对于非负整数 k 有，

$$\begin{aligned}\mathrm{Cov}(C_t, C_{t+1}) &= \delta \mathrm{V}\mathrm{U}\Omega^k\mathrm{U}^{-1}\mathrm{v}' - (\delta \mathrm{v}')^2 \\ &= a\Omega^k b' - a_1 b'_1 = \sum_{i=2}^{m} a_i b'_i \omega_i^k\end{aligned}$$

$$(a_1, a_2, \cdots, a_m)\begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_m \end{pmatrix}^k \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} - a_1 b_1 = \sum_{i=2}^{m} a_i b_i \omega_i^k$$

其中，$a = \delta VU$，$b' = U^{-1}v'$。因此 $\mathrm{Var}(C_t) = \sum_{i=2}^{m} a_i b_i$，并且，对于非负整数 k，

$$\rho(k) \equiv \mathrm{Corr}(C_t, C_{t+1}) = \sum_{i=2}^{m} a_i b_i \omega_i^k \Big/ \sum_{i=2}^{m} a_i b_i$$

这是 ω_2，ω_3，…，ω_m 的 k 次方的加权平均，在某种程度上与高斯（Gaussian）$m-1$ 阶自回归过程相似。以上等式意味着当 $m=2$ 时，对于所有非负整数 k 都有 $\rho(k) = \rho(1)^k$，并且 $\rho(1)$ 是 Γ 的特征值，而 1 不是 Γ 的特征值。

第四节　转移矩阵的估计问题

如果给出一个马尔科夫链的样本数据，并想估计转移概率，有几种方法可以选择，其中一个办法就是找到转移的次数并通过转移次数估计转移概率。例如，具有三个状态马尔科夫链的观测值序列如下：

2121321312　2231213122　1133223222　2321213232　1123322232

3123132121　2233221213　2213233132　3223232131　1132123212

那么转移次数的矩阵为 $(f_{ij}) = \begin{pmatrix} 4 & 12 & 11 \\ 16 & 13 & 14 \\ 7 & 18 & 5 \end{pmatrix}$

其中 f_{ij} 表示观察到的从状态 i 到状态 j 的转移次数。由于从状态 2 到状态 3 的转移次数为 14，并且所有的从状态 2 的转移次数为 16 + 13 + 14，因此 γ_{23} 一个较为可信的估计是 14/43，从而转移矩阵可以估计为：

$$\begin{pmatrix} 4/27 & 12/27 & 11/27 \\ 16/43 & 13/43 & 14/43 \\ 7/30 & 18/30 & 5/30 \end{pmatrix}$$

我们下面证明，这实际上是在给定观察值的条件下求转移概率矩阵 Γ 的最大似然估计。

假设 m 状态马尔科夫链的一个实现（Realization）为 C_1，C_2，…，C_T，我们要从中估计出 m^2-m 个参数 γ_{ij}。以观测值为基础的似然函数为：

$$L=\prod_{i=1}^{m}\prod_{j=1}^{m}\gamma_{ij}^{f_{ij}}$$

求解最大似然函数的过程如下。

首先，我们可以从观察值中找出每一种转移发生的频数矩阵：

$$\begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mm} \end{bmatrix}$$

一般情况下，如果每个事件都是独立的，整个事件发生的可能性就是所有独立事件的概率相乘，于是似然函数计算如下：

$$L=p(x_1)p(x_2)\cdots p(x_m)$$

假设从状态 i 转移到状态 j 的概率为 γ_{ij}，则共发生 f_{ij} 次这样的转移，在写出似然函数时，就应该有 f_{ij} 个 γ_{ij} 相乘，即为 $\gamma_{ij}^{f_{ij}}$。例如可以从状态1、2、…、m 转移到状态 j，转移概率分别为 γ_{1j}、γ_{2j}、…、γ_{mj}。

转移一次后，状态为 j 的概率为：$L_j=\gamma_{1j}^{f_{1j}}\gamma_{2j}^{f_{2j}}\cdots\gamma_{mj}^{f_{mj}}=\prod_{i=1}^{m}\gamma_{ij}^{f_{ij}}$。

所有可能的转移概率相乘得到：

$$L=\prod_{j=1}^{m}L_j=\prod_{j=1}^{m}\prod_{i=1}^{m}\gamma_{ij}^{f_{ij}}$$

对上述似然函数取自然对数，为：

$$\ln L=\sum_{i=1}^{m}\left(\sum_{j=1}^{m}f_{ij}\ln\gamma_{ij}\right)=\sum_{i=1}^{m}l_i$$

我们分别对每一个 l_i 最大化，从而最大化 L。用 $1-\sum_{k\neq i}\gamma_{ik}$ 替换 γ_{ii}，对 l_i 求转移概率矩阵 i 行向量非对角线元素的微分，并使导数等于0。

$$l_i=\sum_{j=1}^{m}f_{ij}\ln\gamma_{ij}=\left(\sum_{j\neq i}f_{ij}\ln\gamma_{ij}\right)+f_{ii}\ln\gamma_{ii}=\left(\sum_{j\neq i}f_{ij}\ln\gamma_{ij}\right)+f_{ii}\ln\left(1-\sum_{j\neq i}\gamma_{ij}\right)$$

上式对 γ_{ij} 求导，可得：

$$\frac{f_{ij}}{\gamma_{ij}} + \frac{-f_{ii}}{1 - \sum_{k\neq i}\gamma_{ik}} = \frac{f_{ij}}{\gamma_{ij}} - \frac{f_{ii}}{\gamma_{ii}} = \frac{\gamma_{ii}f_{ij} - \gamma_{ij}f_{ii}}{\gamma_{ii} * \gamma_{ij}} = 0$$

因此，$f_{ij}\gamma_{ii} = f_{ii}\gamma_{ij}$。

上式两边对 j 求和得：$\sum_{j=1}^{m} f_{ij}\gamma_{ii} = \sum_{j=1}^{m} f_{ii}\gamma_{ij}$。

由于 $\sum_{j=1}^{m} \gamma_{ij} = 1$，从而得到：$\gamma_{ii}\sum_{j=1}^{m} f_{ij} = f_{ii}$。

这表明该似然函数的局部最大化条件为：

$$\gamma_{ii} = \frac{f_{ii}}{\sum_{j=1}^{m} f_{ij}} \qquad \gamma_{ij} = \frac{f_{ij}\gamma_{ii}}{f_{ii}} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$$

因此，这种方法所得到转移概率估计量 $\hat{\gamma}_{ij} = \frac{f_{ii}}{\sum_{k=1}^{m} f_{ik}}$（$i, j = 1, \cdots, m$）可以被视为对 γ_{ij} 的有条件最大似然估计量。转移概率矩阵的估计量还应满足行向量之和等于1的要求。

第二部分

隐蔽马尔科夫模型



第四章

混合分布和隐蔽马尔科夫模型[①]

隐蔽马尔科夫模型（Hidden Markov Model，简称“HMM”）是指所产生观测值的分布取决于潜在状态的模型。这种状态变量具有两个特点：首先，该状态变量是不可观测的；其次，它服从马尔科夫过程。HMM 在时间序列分析中得到广泛应用，尤其是在对离散型时间序列的处理中。

HMM 被用于信号处理已经有 30 多年的历史，在自动语音识别等方面得到了非常成功的应用。目前，这一理论的研究和应用已经延伸到了很多其他领域：特征识别系统方面，包括人脸识别、姿态手势识别、字迹签名识别等；环境科学方面，包括风向、降雨量、地震预测等；金融投资方面，如每日收益时间序列的分析；等等。例如，研究金融市场收益率时间序列 x_t 的分布。根据市场收益率发展形态将市场分为牛市和熊市两类，相应地，观测值也就被分为两组。在每个分组中市场收益率的分布可能都服从正态分布，但所有时间范围内的收益率 x_t 就不一定服从正态分布了。也就是说，x_t 可能只是局部服从正态分布，而整个分布也许需要几个分布组成的混合分布模型来表示。其中，不可观测的、代表收益率所属牛市或熊市组别的变量即为状态变量，该状态变量服从马尔科夫过程，且该状态变量所决定的个别正态分布生成了 HMM 的观测值 x_t 。

HMM 的主要优点就在于所使用算法的简洁性和通用性，尤其是对其中的参数可

---

① 尽管我们使用“隐蔽马尔科夫模型”一词，但它并不是描述这些模型或类似模型的唯一名字。例如，人们也可能经常使用以下名称来描述此模型：“隐蔽马尔科夫过程”“独立的马尔科夫混合模型”“马尔科夫转换模型”“服从马尔科夫状态转换的模型”或“马尔科夫混合模型”。

以直接求解。

本章将对 HMM 及其用途进行简要的介绍。首先从最简单的情况入手，对状态序列相互独立条件下的混合分布模型及其参数估计进行分析。

第一节　状态序列相互独立的混合分布模型

对于大多数的样本数据而言，我们都很难使用一个简单的、标准的泊松分布来进行描述。这是因为泊松分布的分布函数是 $p(x) = e^{-\lambda}\lambda^{x}/x!$，它的均值和方差都是 λ。举例如图 4－1，假设样本数据的方差 $s^2 \approx 50$，均值 $\bar{x} \approx 20$，方差比均值大很多，只用一个泊松分布来表示数据的生成过程并不合适。图中各点描述的是拟合的泊松分布，各竖线表示本例中实际数据的分布。由图可见，两个分布相差较大，本例中数据的分布并不能只用一个简单的泊松分布来表示。

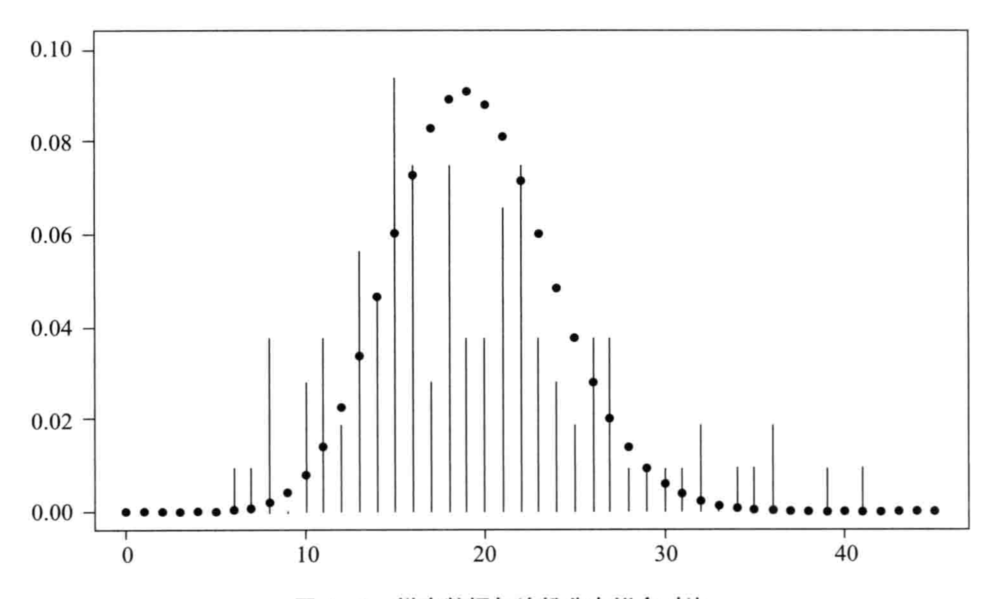


图 4－1　样本数据与泊松分布拟合对比

但是，我们可以把几个分布组合在一起，形成一个混合分布模型，并用该混合分布模型来解决上述观测值过于离散的问题。建立混合分布模型能够把数据中观测不到的异质性考虑进来。比如说，将数据分成差异明显的若干组，每组观测值分别服从一个特定的分布。

按照这样的思路将股市收益率数据通过两个泊松分布来表示。这两个分布的均值分别为 λ_1 和 λ_2 ，何时取 λ_1 或 λ_2 是由另外一个随机过程来决定的。假设取 λ_1 的概率为 w_1 ，取 λ_2 的概率为 $w_2 = 1 - w_1$ 。后面会证明这个混合分布模型的方差比均值大，二者的差为 $w_1 w_2 (\lambda_1 - \lambda_2)^2$ 。

如果一个混合分布由 m 个子分布组成，那么混合分布就是这些单个分布的线性组合，这些单个分布可能是离散或连续的。在由两个分布组成的混合分布中，这个混合分布取决于两个概率分布或概率密度函数 $p_1(x)$ 和 $p_2(x)$ 。

为了清楚地表示混合分布的构成，我们需要一个离散的随机变量 S 来表示单个分布在混合分布中的比重：

$$S = \begin{cases} 1 & \text{选择分布 1 的概率为 } w_1 \\ 2 & \text{选择分布 2 的概率为 } w_2 \end{cases}$$

由这两个相互独立的子分布组成混合分布，假设 X 表示服从该混合分布的随机变量，则 X 的分布函数为 $X = p_1(x) \cdot w_1 + p_2(x) \cdot w_2$ ；X 的期望值和方差分别为：

$$\mathrm{E}(X) = w_1 \mathrm{E}(X_1) + w_2 \mathrm{E}(X_2) = w_1 \lambda_1 + w_2 \lambda_2$$

$$\mathrm{Var}(X) = w_1 \lambda_1^2 + w_2 \lambda_2^2 + w_1 w_2 (\lambda_1 - \lambda_2)^2$$

但 S 的值是不确定的，而且随着时间的变化而变化，也就是说我们不知道在某个时点上混合分布应该取两个分布中的哪一个。图 4－2 展示了由两个分布组成的混合分布的结构。此例中，S_t 的值表示熊市和牛市两种状态，在熊市时 $S_t = 1$ ，此时观测值服从分布 $p_1(x)$ ；在牛市时 $S_t = 2$ ，观测值服从分布 $p_2(x)$ 。

很容易将上述思路推广到混合分布的子分布个数扩大到 m 的情况。令 $w_1, \cdots, w_m$ 分别代表每个子分布被选出的概率，$p_1, \cdots, p_m$ 代表各自子分布的概率分布，X 表示由

这 m 个分布组成的混合分布的随机变量。

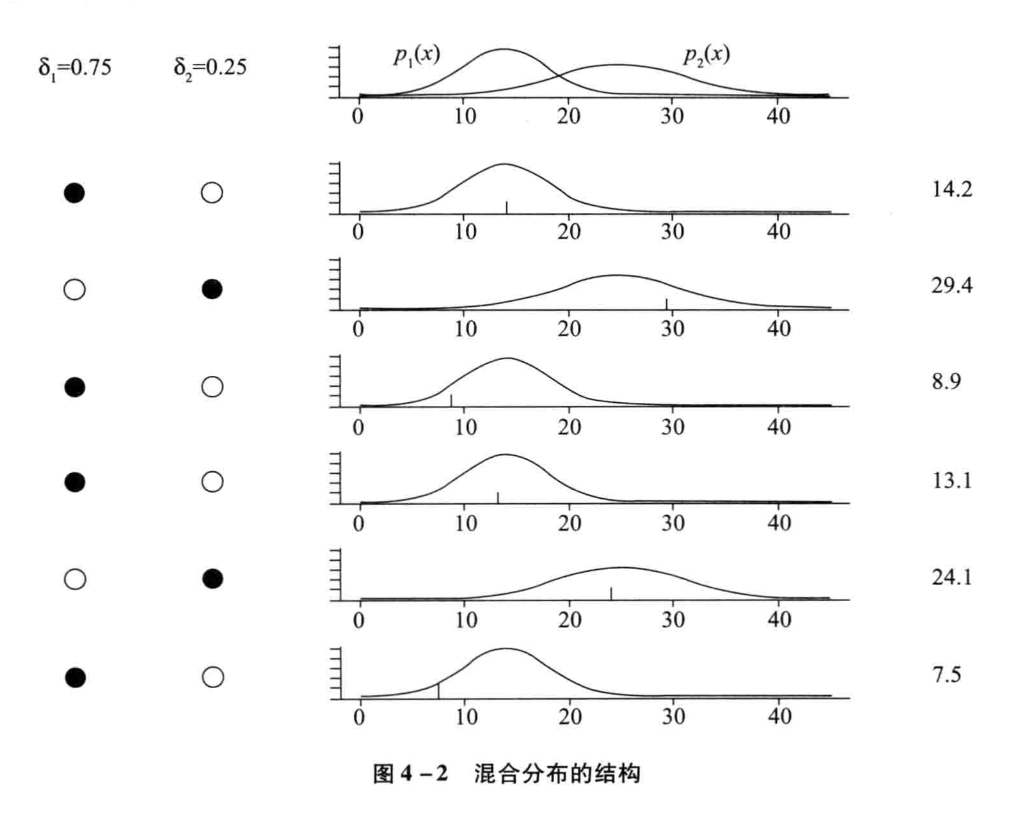


图 4－2　混合分布的结构

如果 X 是连续随机变量，则 X 的分布函数为 $p(x) = \sum_{i=1}^{m} \delta_i p_i(x)$；如果 X 是离散随机变量，则 X 的分布函数为 $p(X = x) = \sum_{i=1}^{m} p(X = x \mid S = i) p(S = i)$。

服从混合分布的变量的期望可以由各个组成分布的期望表示出来，其中 Y_i 表示服从分布 p_i 的随机变量：

$$\mathrm{E}(x) = \sum_{i=1}^{m} p(S = i)\mathrm{E}(X \mid S = i) = \sum_{i=1}^{m} w_i \mathrm{E}(Y_i) \tag{4.1}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = w_1 \mathrm{Var}(Y_1) + w_2 \mathrm{Var}(Y_2) + w_1 w_2 (\mathrm{E}(Y_1) - \mathrm{E}(Y_2))^2 \tag{4.2}$$

上式证明如下：

$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$

$$= w_1 \mathrm{E}(Y_1^2) + w_2 \mathrm{E}(Y_2^2) - [w_1 \mathrm{E}(Y_1) + w_2 \mathrm{E}(Y_2)]^2$$

$$= w_1(\mathrm{E}(Y_1^2) - \mathrm{E}(Y_1)^2) + w_2(\mathrm{E}(Y_2^2) - \mathrm{E}(Y_2)^2) + w_1 \mathrm{E}(Y_1)^2 + w_2 \mathrm{E}(Y_2)^2$$
$$- w_1{}^2 \mathrm{E}(Y_1)^2 - w_2{}^2 \mathrm{E}(Y_2)^2 - 2w_1 w_2 \mathrm{E}(Y_1)\mathrm{E}(Y_2)$$

$$= w_1 \mathrm{Var}(Y_1) + w_2 \mathrm{Var}(Y_2) + w_1(1 - w_1)\mathrm{E}(Y_1)^2 + w_2(1 - w_2)\mathrm{E}(Y_2)^2$$
$$- 2w_1 w_2 \mathrm{E}(Y_1)\mathrm{E}(Y_2)$$

$$= w_1 \mathrm{Var}(Y_1) + w_2 \mathrm{Var}(Y_2) + w_1 w_2 \mathrm{E}(Y_1)^2 + w_2 w_1 \mathrm{E}(Y_2)^2$$
$$- 2w_1 w_2 \mathrm{E}(Y_1)\mathrm{E}(Y_2)$$

$$= w_1 \mathrm{Var}(Y_1) + w_2 \mathrm{Var}(Y_2) + w_1 w_2 (\mathrm{E}(Y_1) - \mathrm{E}(Y_2))^2$$

第二节 状态相互独立混合分布的参数估计

通常用极大似然法来估计混合分布中的参数。对于一个由 m 个分布组成的混合分布，无论它是离散的还是连续的，都有：

$$L(\theta_1, \cdots, \theta_m, w_1, \cdots, w_m \mid x_1, \cdots, x_n) = \prod_{j=1}^{n} \sum_{i=1}^{m} w_i p_i(x_j, \theta_i) \tag{4.3}$$

其中，$\theta_1, \cdots, \theta_m$ 分别为组成混合分布的 m 个分布的分布参数；$w_1, \cdots, w_m$ 分别为 m 个分布被取到的概率，其和为1；$x_1, \cdots, x_n$ 代表 n 个样本观测值。如果这 m 个分布都分别只含有一个参数，那么共有 $2m-1$ 个待估参数，包括 m 个 θ 和 $m-1$ 个 w 。

但上式并不容易用极大似然法来予以估计。比如，假设混合分布由两个独立的泊松分布组成，均值分别为 λ_1 和 λ_2 ，被取到的概率为 w_1 和 w_2 ，则混合分布的分布函数为 $p(x) = w_1 \dfrac{\lambda_1^x e^{-\lambda_1}}{x!} + w_2 \dfrac{\lambda_2^x e^{-\lambda_2}}{x!} = w_1 \dfrac{\lambda_1^x e^{-\lambda_1}}{x!} + (1 - w_1)\dfrac{\lambda_2^x e^{-\lambda_2}}{x!}$。

此时，只有 λ_1 、λ_2 和 w_1 三个待估参数。极大似然函数为：

$$L(\lambda_1, \lambda_2, w_1 \mid x_1, \cdots, x_n) = \prod_{i=1}^{n} (w_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + (1 - w_1)\frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!})$$

极大似然函数 $L(\lambda_1, \lambda_2, w_1 \mid x_1, \cdots, x_n)$ 由 n 项的乘积组成，而每一项都是和的形

式，所以等式两边取对数并不能使问题得到简化。此时，有两种方法来解决这个问题：一种是数值解法，另一种是 EM 法。

第三节　简单隐蔽马尔科夫模型

在前面的分析中，我们假设每一个观测值都是由 m 个泊松分布中的一个随机产生，这 m 个泊松分布的均值分别是 $\lambda_1,\lambda_2,\cdots,\lambda_m$ ，而均值 λ_i 发生的概率为 $w_i(i=1,2,\cdots,m)$ ，并且 $\sum_{i=1}^{m} w_i = 1$ 。以此建立的混合模型解决了方差过度分散的问题，但是这种简单的状态序列相互独立的混合模型并不能够解决观测值之间相互依赖的情况，放宽独立性假设是解决序列自相关的一个办法。为了简化问题，人们通常都假设状态服从马尔科夫过程，这种模型就叫做**泊松－隐蔽马尔科夫模型**。

假设分布取决于从第 1 期到第 t 期不可观测的状态变量序列 S_1^t ，并根据 S_1^t 所确定的分布来产生 X_1^t 。隐蔽马尔科夫模型可以表示为：

$$P(S_t \mid S_1^{t-1}) = P(S_t \mid S_{t-1}) \qquad t = 2,3,\cdots \tag{4.4}$$

$$P(X_t \mid X_1^{t-1}, S_1^t) = P(X_t \mid S_t) \qquad t \in N \tag{4.5}$$

该模型有由两部分组成：一部分是不能被观测到的状态过程 $\{S_t : t = 1,2,\cdots\}$ ，该过程满足马尔科夫性；另一部分是状态依赖变量 $\{X_t : t = 1,2,\cdots\}$ 。第二个式子表示 X_t 依赖于状态变量 S_t 。当 S_t 已知时，X_t 的分布只依赖于当前的状态 S_t ，与之前的状态 S_1^{t-1} 以及观测值 X_1^{t-1} 无关。

图 5－3 展示了隐蔽马尔科夫模型产生随机观测值的过程，其中取决于状态的子分布为 p_1 和 p_2 ，这两个分布 p_1 和 p_2 被分配到的概率为 $w=(0.75, 0.25)$ ，同时状态转换矩阵 $\Gamma = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$ 。与独立混合情形相比，这里 S_t 不再独立于 S_{t-1} ，而是依赖于 S_{t-1} 。

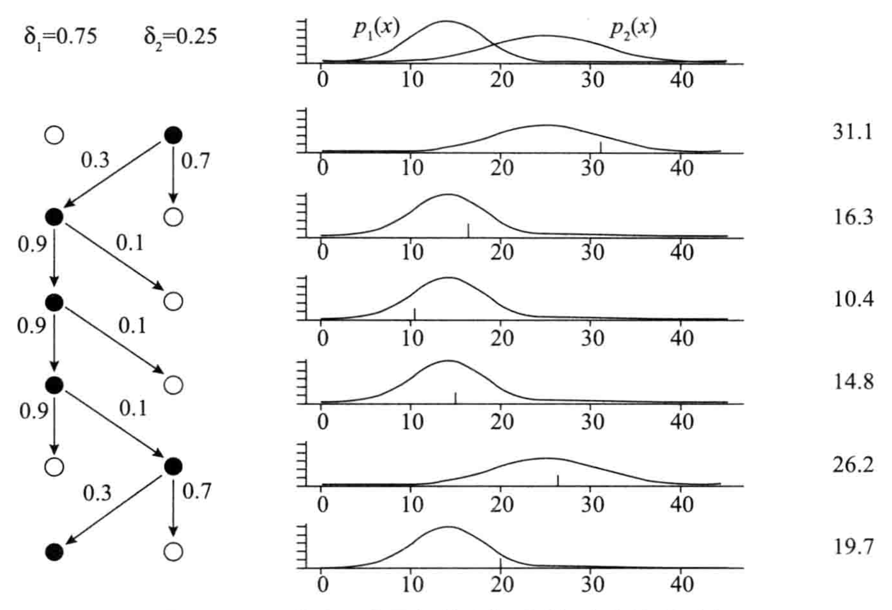


图 4－3　两状态隐蔽马尔科夫模型随机产生的观测值

注：如图左侧所示，这个隐蔽马尔科夫链按照状态 2，1，1，1，2，1 的路径演化，图中间表示了给定状态下的分布，图右侧的观测值由状态所决定的分布产生。

下面我们考虑离散的情况，定义如下变量：

$$p_i(x) = p(X_t = x \mid S_t = i)\ ,\ i = 1,2,\cdots,m$$

其中，p_i 指当马尔科夫链 X_t 在 t 期、状态 i 下的概率分布函数。连续的情况也是相似的，定义 p_i 为马尔科夫链 X_t 在 t 期、状态 i 下的概率密度函数。HMM 状态依赖分布模型就是指 m 个单独分布 p_i 组成的混合模型。

为了方便起见，我们仅仅给出离散的状态依赖型分布的结论，连续情形下的结论也可以用类似的方法得到。对于离散型变量 $X_t(t = 1,2,\cdots,T)$ ，定义 $u_i(t) = P(S_t = i)$ ，则：

$$P(X_t) = \sum_{i=1}^{m} P(X_t = x \mid S_t = i)P(S_t = i) = \sum_{i=1}^{m} p_i(x)u_i(t) \qquad (4.6)$$

该式可以被写成矩阵的形式如下：

$$P(X_t) = (u_1(t),\cdots,u_m(t))\begin{pmatrix} p_1(x) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_m(x) \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = u(t)\mathrm{P}(x)1' \quad (4.7)$$

其中，$\mathrm{P}(x)$ 定义为对角线元素为 $p_i(x)$ 的对角矩阵。从前面我们可知 $u(t) = u(1)\Gamma^{t-1}$，因此可得：$P(X_t) = u(1)\Gamma^{t-1}\mathrm{P}(x)1'$。

上式要求当马尔科夫链是齐次的，但不要求稳态性。如果我们假设马尔科夫链是稳态的，且其稳态分布为 δ，上述表达形式将会更加简单：

对于任意的 $t \in N$，$\delta\Gamma^{t-1} = \delta$，因此 $P(X_t) = \delta\mathrm{P}(x)1'$。 (4.8)

如前所述，与 HMM 相关的多变量分布计算很简单。在任何给定的模型下，一组随机变量 V_i 的联合分布如下所示：

$$P(V_1,V_2,\cdots,V_n) = \prod_{i=1}^{n} P(V_i \mid pa(V_i)) \quad (4.9)$$

其中 $pa(V_i)$ 为 V_i 的父辈（Parent），这一图论中的概念表示 V_i 所依赖的信息集。

对于正整数 k，考虑 X_t,X_{t+k},S_t,S_{t+k} 这四个随机变量的逻辑关系，$pa(S_t)$ 是空集，$pa(X_t) = \{S_t\}$，$pa(S_{t+k}) = \{S_t\}$，$pa(X_{t+k}) = \{S_{t+k}\}$。因此可得到下式：

$$P(X_t,X_{t+k},S_t,S_{t+k}) = P(S_t)P(X_t \mid S_t)P(S_{t+k} \mid S_t)P(X_{t+k} \mid S_{t+k}) \quad (4.10)$$

所以有：

$$\begin{aligned} P(X_t = v,X_{t+k} = w) &= \sum_{i=1}^{m}\sum_{j=1}^{m} P(X_t = v,X_{t+k} = w,S_t = i,S_{t+k} = j) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} P(S_t = i)p_i(v)P(S_{t+k} = j \mid S_t = i)p_j(w) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t)p_i(v)\Gamma_{ij}(k)p_j(w) \end{aligned} \quad (4.11)$$

将上述求和过程写成矩阵乘积形式：

$$P(X_t = v,X_{t+k} = w) = u(t)\mathrm{P}(v)\Gamma^k\mathrm{P}(w)1' \quad (4.12)$$

如果马尔科夫链是稳态的，则上式可化简为：

$$P(X_t = v,X_{t+k} = w) = \delta\mathrm{P}(v)\Gamma^k\mathrm{P}(w)1' \quad (4.13)$$

在稳态马尔科夫链的情况下，三变量分布的公式如下所示：

$$P(X_t = v, X_{t+k} = w, X_{t+k+l} = z) = \delta \mathrm{P}(v)\Gamma^k \mathrm{P}(w)\Gamma^l \mathrm{P}(z)1' \tag{4.14}$$

由上述结果可知：

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \mathrm{E}(X_t \mid S_t = i)P(S_t = i) = \sum_{i=1}^{m} u_i(t)\mathrm{E}(X_t \mid S_t = i) \tag{4.15}$$

在稳态条件下，上式可以化简为：

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \delta_i \mathrm{E}(X_t \mid S_t = i) \tag{4.16}$$

由上述公式可以推导出处于稳态条件下两状态泊松－隐蔽马尔科夫模型的计算公式为：

$$\mathrm{E}(X_t) = \delta_1 \lambda_1 + \delta_2 \lambda_2$$

在一般情况下，对于状态存在相关性的任意函数 g 的期望，关于 $E(g(X_t))$ 和 $E(g(X_t, X_{t+k}))$ 的类似结论依然存在。在稳态马尔科夫链条件下：

$$\mathrm{E}(g(X_t)) = \sum_{i=1}^{m} \delta_i \mathrm{E}(g(X_t) \mid C_t = i) \tag{4.17}$$

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i=1}^{m} \mathrm{E}(g(X_t) \mid C_t = i)\delta_i \Gamma_{ij}(k) \tag{4.18}$$

其中，$\Gamma_{ij}(k) = (\Gamma^k)_{ij}$，$k \in N$。通常，对于能拆分成 $g(X_t, X_{t+k}) = g_1(X_t)g_2(X_{t+k})$ 的函数 g，上式等价于：

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g_1(X_t) \mid C_t = i)\mathrm{E}(g_2(X_{t+k}) \mid C_{t+k} = j)\delta_i \Gamma_{ij}(k) \tag{4.19}$$

由上述公式可以推导出协方差、相关系数的计算公式。比如，稳态条件下两状态泊松－隐蔽马尔科夫模型的计算公式为：

$$\mathrm{Var}(X_t) = \mathrm{E}(X_t) + \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 > \mathrm{E}(X_t)$$

$$\mathrm{Cov}(X_t, X_{t+k}) = \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 (1 - \Gamma_{12} - \Gamma_{21})^k$$

注意：这里 X_t 和 X_{t+k} 的协方差公式是 $\rho(k) = A(1 - \Gamma_{12} - \Gamma_{21})^k$ 的形式，其中 $A \in [0,1)$，且当 $\lambda_1 = \lambda_2$ 时，$A = 0$。

下面，针对稳态条件下两状态泊松－隐蔽马尔科夫模型，我们来推导混合方差

的公式 $\mathrm{Var}(X_t) = \mathrm{E}(X_t) + \delta_1\delta_2(\lambda_2 - \lambda_1)^2$。

具体推导过程如下：

$X_t = \delta_1 X_1 + \delta_2 X_2$

$\mathrm{E}(X_t) = \delta_1\lambda_1 + \delta_2\lambda_2$

$g(x_t) = [X_t - \mathrm{E}(X_t)]^2 = [X_t - (\delta_1\lambda_1 + \delta_2\lambda_2)]^2$

$\{g(x_t) \mid C_t = 1\}$

$= [X_{1t} - \mathrm{E}(X_t)]^2$

$= [X_{1t} - (\delta_1\lambda_1 + \delta_2\lambda_2)]^2$

$= [(X_{1t} - \lambda_1 + (1-\delta_1)\lambda_1 - \delta_2\lambda_2)]^2$

$= [(X_{1t} - \lambda_1) + \delta_2(\lambda_1 - \lambda_2)]^2$

$= (X_{1t} - \lambda_1)^2 + 2(X_{1t} - \lambda_1)\delta_2(\lambda_1 - \lambda_2) + \delta_2^2(\lambda_1 - \lambda_2)^2$

$\mathrm{E}\{g(x_t) \mid C_t = 1\}$

$= \mathrm{E}\{(X_{1t} - \lambda_1)^2 + 2(X_{1t} - \lambda_1)\delta_2(\lambda_1 - \lambda_2) + \delta_2^2(\lambda_1 - \lambda_2)^2\}$

$= \mathrm{E}(X_{1t} - \lambda_1)^2 + 2\mathrm{E}\{(X_{1t} - \lambda_1)\delta_2(\lambda_1 - \lambda_2)\} + \delta_2^2(\lambda_1 - \lambda_2)^2$

$= (X_{1t} - \lambda_1)^2 + \delta_2^2(\lambda_1 - \lambda_2)^2$

$= \lambda_1 + \delta_2^2(\lambda_1 - \lambda_2)^2$

类似可得，$\mathrm{E}\{g(x_t) \mid C_t = 2\} = \lambda_2 + \delta_1^2(\lambda_2 - \lambda_1)^2$。

所以 $\mathrm{Var}(X_t) = \mathrm{E}(g(X_t)) = \sum_{i=1}^{2}\delta_i \mathrm{E}(g(X_t) \mid C_t = i)$

$= \delta_1\mathrm{E}\{g(x_t) \mid C_t = 1\} + \delta_2\mathrm{E}\{g(x_t) \mid C_t = 2\}$

$= \delta_1[\lambda_1 + \delta_2^2(\lambda_1 - \lambda_2)^2] + \delta_2[\lambda_2 + \delta_1^2(\lambda_2 - \lambda_1)^2]$

$= (\delta_1\lambda_1 + \delta_2\lambda_2) + \delta_1\delta_2^2(\lambda_1 - \lambda_2)^2 + \delta_2\delta_1^2(\lambda_2 - \lambda_1)^2$

$= \mathrm{E}(X_t) + \delta_1\delta_2(\lambda_1 - \lambda_2)^2$

第四节　隐蔽马尔科夫模型的极大似然函数

假设隐蔽马尔科夫模型有 m 个状态，并产生 T 个连续的观测值 $x_1, x_2, \cdots, x_T$，本节将讨论如何求得该观测值序列的似然函数 L_T。从直观来讲，这个似然函数的计算包含 m^T 个和项，其中每一项是 $2T$ 个因子的乘积，所以需要处理 Tm^T 的同阶量次的运算问题，而实践证明这样直接计算似然函数几乎是不可行的。Baum（1972）等人研究证明用迭代的办法能够使上述似然函数的计算可行。

如果似然函数可以用简单的形式表达，我们就能够通过最大化似然函数来估计参数。下面，我们来说明似然函数 L_T 相对于 m^2T 同阶量次运算是可计算的。

首先，我们探究两状态隐蔽马尔科夫模型的似然函数，然后再将其推广到更为一般的情况。考虑具有如下转换矩阵的两状态隐蔽马尔科夫模型：

$$\Gamma = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

且给定状态下的分布函数为：

$$P(X_t = x \mid S_t = 1) = \frac{1}{2}, x = 0 \text{ 或 } 1$$

$$P(X_t = 1 \mid S_t = 2) = 1$$

我们将上述模型称为**伯努利隐蔽马尔科夫模型**。在该例子中马尔科夫链的均衡状态为 $\delta = \frac{1}{7}(3,4)$，则 $X_1 = X_2 = X_3 = 1$ 的概率可以被写成如下形式：

$$P(X_1, X_2, X_3, S_1, S_2, S_3) = P(S_1)P(X_1 \mid S_1)P(S_2 \mid S_1)P(X_2 \mid S_2)P(S_3 \mid S_2)P(X_3 \mid S_3) \tag{4.20}$$

对 S_1, S_2, S_3 求和可得：

$$P(X_1=1,X_2=1,X_3=1)=\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}P(X_1=1,X_2=1,X_3=1,S_1=i,S_2=j,S_3=k)$$

$$=\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}w_ip_i(1)\Gamma_{ij}p_j(1)\Gamma_{jk}p_k(1) \tag{4.21}$$

从上式可知，有 $m^T=2^3$ 项，其中每一项都是 $2T=2\times3$ 项的乘积。

用矩阵表示上述和的形式将会更加简便。定义 $\mathrm{P}(u)$ 为对角线元素为 $(p_1(u),p_2(u))$ 的对角矩阵，即：

$$\mathrm{P}(0)=\begin{pmatrix}\frac{1}{2}&0\\0&0\end{pmatrix}\quad \mathrm{P}(1)=\begin{pmatrix}\frac{1}{2}&0\\0&1\end{pmatrix}$$

从而上式可以被表达为：

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}w_ip_i(1)\Gamma_{ij}p_j(1)\Gamma_{jk}p_k(1)=w\mathrm{P}(1)\Gamma\mathrm{P}(1)\Gamma\mathrm{P}(1)1' \tag{4.22}$$

下面我们来讨论隐蔽马尔科夫模型似然函数的一般形式。

假设隐蔽马尔科夫模型有 m 个状态，初始分布为 w，转换矩阵为 Γ，观测值在给定状态 i 条件下的概率密度函数为 p_i 的，生成的一列观察值为 $x_1,x_2,\cdots,x_T$，我们的目标是求得产生该序列的概率 L_T。

首先证明隐蔽马尔科夫模型似然函数是如下形式：

$$L_T=w\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma\mathrm{P}(x_3)\cdots\Gamma\mathrm{P}(x_T)1' \tag{4.23}$$

如果 S_1 的分布 w 是稳态马尔科夫链的稳态分布 δ，那么：

$$L_T=\delta\Gamma\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma\mathrm{P}(x_3)\cdots\Gamma\mathrm{P}(x_T)1' \tag{4.24}$$

在证明上述问题之前，我们需要定义一个新的矩阵 $B_t=\Gamma\mathrm{P}(x_t)$ 来重新表述这个问题。所以以上两式可以分别被写作：

$$L_T=w\mathrm{P}(x_1)\mathrm{B}_2\mathrm{B}_3\cdots\mathrm{B}_T1' \tag{4.25}$$

$$L_T=\delta B_1\mathrm{B}_2\mathrm{B}_3\cdots\mathrm{B}_T1' \tag{4.26}$$

具体的证明过程如下（仅介绍离散情况）：

首先，$L_T=P(X^{(T)}=x^{(T)})=\sum_{c_1,c_2,\cdots,c_T=1}^{m}P(X^{(T)}=x^{(T)},S^{(T)}=s^{(T)})$。

通过前面的分析，我们可以得到：

$$P(X^{(T)},S^{(T)}) = P(S_1)\prod_{k=2}^{T}P(S_k \mid S_{k-1})\prod_{k=1}^{T}P(X_K \mid S_k)$$

并得出：

$$\begin{aligned} L_T &= \sum_{S_1,S_2,\cdots,S_T=1}^{m}(w_{S_1}\gamma_{S_1,S_2}\gamma_{S_2,S_3}\cdots\gamma_{S_{T-1},S_T})(p_{S_1}(x_1)p_{S_2}(x_2)\cdots p_{S_T}(x_T)) \\ &= \sum_{S_1,S_2,\cdots,S_T=1}^{m}w_{S_1}p_{S_1}(x_1)\gamma_{S_1,S_2}p_{S_2}(x_2)\gamma_{S_2,S_3}\cdots\gamma_{S_{T-1},S_T}p_{S_T}(x_T) \\ &= w\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma\mathrm{P}(x_3)\cdots\Gamma\mathrm{P}(x_T)1' \end{aligned}$$

如果 w 是稳态马尔科夫链的稳态分布 δ，则有 $\delta\mathrm{P}(x_1) = \delta\Gamma\mathrm{P}(x_1) = \delta\mathrm{B}_1$，因此可知，上式虽然多了一项 Γ，但几乎没有影响。

为了给出似然函数计算方法，我们来定义向量 f_t：

$$f_t = w\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma\mathrm{P}(x_3)\cdots\Gamma\mathrm{P}(x_t) = w\mathrm{P}(x_1)\prod_{s=2}^{t}\Gamma\mathrm{P}(x_s)\ ,\ t = 1,2,\cdots,T \tag{4.27}$$

由此，我们则可以得出如下方程：

$$L_T = f_T1'\text{，且当 } t \geq 2 \text{ 时有 } f_t = f_{t-1}\Gamma\mathrm{P}(x_t)\text{。} \tag{4.28}$$

基于以上表达方式，我们可以方便地对似然函数进行求解，具体算法如下：

当 $t = 1$ 时，$f_1 = w\mathrm{P}(x_1)$。

当 $t = 2,3\cdots,T$ 时，$f_t = f_{t-1}\Gamma\mathrm{P}(x_t)$。

通过迭代最终得到 $L_T = f_T1'$。

从以上算法可知，向量 f_{t-1} 与状态转移矩阵 Γ 的乘积共运算 m 次，$f_{t-1}\Gamma$ 再与 $P(x_t)$ 的乘积共 m 次。因此，两者共运算 m^2 次。由于我们的算法是由 $t = 1,2,\cdots,T$ 的递归算法，所以求解似然函数的运算次数是 Tm^2 次。换句话说，对于递归循环中的每一个 t，我们需要计算向量 f_{t-1} 与状态转移矩阵 Γ 的乘积，这需要运算 m 次；然后，再用此结果 $(f_{t-1}\Gamma)$ 中的 m 个元素去乘以状态依赖概率向量 $\mathrm{P}(x_t)$，这样就需要 运算 $m \times m$ 次。因此，共需要计算 Tm^2。

最后，我们来讨论数据缺失的情况下如何对似然函数进行调整的问题。在 HMM 中，可以通过对似然函数进行简单调整来处理数据缺失问题。假设一个 HMM 有观

测值 $x_1,x_2,x_4,x_7,x_8,\cdots,x_T$，但是 x_3,x_5,x_6 这几个数据是缺失的，则有似然函数：

$$P(X_1 = x_1,X_2 = x_2,X_4 = x_4,X_7 = x_7,\cdots,X_T = x_T)$$

$$= \sum w_{S_1}\gamma_{S_1,S_2}\gamma_{S_2,S_4}(2)\gamma_{S_4,S_7}(3)\gamma_{S_7,S_8}\cdots\gamma_{S_{T-1},S_T} \times p_{S_1}(x_1)p_{S_2}(x_2)p_{S_4}(x_4)p_{S_7}(x_7)\cdots p_{S_T}(x_T) \tag{4.29}$$

其中 $\gamma_{ij}(k)$ 表示 k 阶转换概率，总和包括除了 S_3,S_5,S_6 以外的所有的 S_t。因为

$$\sum w_{S_1}p_{S_1}(x_1)\gamma_{S_1,S_2}p_{S_2}(x_2)\gamma_{S_2,S_4}(2)p_{S_4}(x_4)\gamma_{S_4,S_7}(3)p_{S_7}(x_7)\cdots\gamma_{S_{T-1},S_T}p_{S_T}(x_T)$$

$$= w\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma^2\mathrm{P}(x_4)\Gamma^3\mathrm{P}(x_7)\cdots\Gamma\mathrm{P}(x_T)1' \tag{4.30}$$

用 $L_T^{-(3,5,6)}$ 代表除了 x_3,x_5,x_6 以外的所有观测值的似然函数，所以：

$$L_T^{-(3,5,6)} = \mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\Gamma^2\mathrm{P}(x_4)\Gamma^3\mathrm{P}(x_7)\cdots\Gamma\mathrm{P}(x_T)1' \tag{4.31}$$

这个结论意味着，似然函数中与缺失变量相对应的对角矩阵 $\mathrm{P}(x_t)$ 被单位矩阵所替换。也就是说，相应的概率 $p_i(x_t)$ 在所有状态中全部用 1 代替了。由此可见，即使存在缺失数据，HMM 的似然函数也可以很容易计算，这在条件分布的推导中非常有用。

假设在隐蔽马尔科夫模型的一系列观测值中，有部分观测值是区间型的。比如，可能仅仅能确定当 $4 \le t \le T$ 时 x_t 的真实值，以及 $x_1 \le 15$，$20 \le x_2 \le 30$，$x_3 > 100$。在这种情况下，可以将似然函数中的对角矩阵 $P(x_1)$ 替换为矩阵：

$$\begin{bmatrix} P(X_1 \le 15 \mid S_1 = 1) & 0 \\ 0 & P(X_1 \le 15 \mid S_1 = 2) \end{bmatrix}$$

同理，可替换 $\mathrm{P}(x_2)$ 和 $\mathrm{P}(x_3)$。更一般地，假设 $c \le x_t \le d$，马尔科夫链可能有 m 个状态。可以通过用 $m \times m$ 阶对角线元素为 $P(c \le x_t \le d \mid S_t = i)$ 的对角矩阵替换似然函数中的 $\mathrm{P}(x_t)$。对于缺省值，也可以将其视为区间型变量来进行处理。



第五章

隐蔽马尔科夫模型极大似然函数估计方法

从前面的分析中可以知道，如果马尔科夫链是稳态的，那么对于稳态分布 δ 来说有 $\delta = \delta T$ 。对于 $t = 1,2,\cdots,T$ 我们有以下递归算法：

$$f_0 = \delta$$

$$f_t = f_{t-1}\Gamma\mathrm{P}(x_t) \tag{5.1}$$

根据上一章内容，在平稳条件下共需要 Tm^2 次运算即可求得似然函数。因此，即使 T 很大，对似然函数的估计也是可行的。这样，我们就可以通过数值算法来解似然函数最大化问题，并在此基础上直接估计参数值。

但是，直接使用这种数值算法会遇到一些问题。主要问题包括：数值下溢、参数取值范围约束、似然函数极值点不唯一，等等。本章将首先讨论怎样克服这些问题，以便使用简便的递归数值算法去估计极大似然函数值；然后，在引入前向和后向概率的基础上，讨论参数估计的 EM 算法。

第一节　数值算法

在观测值离散的情况下，f_t 中的元素可以表达成概率乘积的形式，并随着 t 的增加不断减小，逐渐接近于0。鉴于 f_t 与似然值的关系，似然函数可能以指数速度趋向于0或者无穷。这种数值溢出并不仅局限于离散分布的下溢问题，也可能会出现于连续分布的上溢问题。无论上溢还是下溢，处理的方式并无太大区别，因此我们仅以下溢为例来展开讨论。

似然函数是转移矩阵元素与向量乘积之和，HMM似然函数的计算要比独立混合模型似然函数计算复杂得多，因此不能仅仅依靠对似然函数取自然对数的方法来避免数值下溢问题。为了解决这个问题，Durbin 等人（1998）提出了基于以下近似计算的一种方法：

假设 $u > v$，我们希望计算 $\ln(u + v)$，那么有：

$$\ln u + \ln(1 + v/u) = \ln u + \ln(1 + \exp(\tilde{v} - \tilde{u}))$$

其中 $\tilde{u} = \ln u$，$\tilde{v} = \ln v$。$\ln(1 + e^x)$ 可通过插值法估计，并通过参照取值表把估计精度提高到一个合理的范围。

下面通过对前向概率 f_t 进行加权的方法来计算 L_T。

对于 $t = 0,1,\cdots,T$，定义加权前向概率向量为：

$$\alpha_t = f_t/\rho_t \tag{5.2}$$

其中，$\rho_t = \sum_i f_t(i) = f_t 1'$ (5.3)

首先，由 α_t 和 ρ_t 的定义可以直接得到：

$$f_t = (f_t/\rho_t)\rho_t = \alpha_t\rho_t \tag{5.4}$$

$$f_{t-1} = \alpha_{t-1}\rho_{t-1}$$

因此，$f_t = f_{t-1}\Gamma P(x_t)$ 可以写作：

$$\alpha_t\rho_t = \alpha_{t-1}\rho_{t-1}\mathrm{B}_t \tag{5.5}$$

这样，我们的算法由 δ 计算得到 ρ_0 和 α_0 ，并通过不断迭代求得 L_T ：

$$\rho_0 = f_0 1' = \delta 1' = 1$$

$$\alpha_0 = \delta$$

$$\rho_t\alpha_t = \rho_{t-1}\alpha_{t-1}\mathrm{B}_t$$

$$L_T = \rho_T(\alpha_T 1') = \rho_T$$

因此，$L_T = \rho_T = \prod_{t=1}^{T}(\rho_t/\rho_{t-1})$。 (5.6)

将 $\rho_t\alpha_t = \rho_{t-1}\alpha_{t-1}B_t$ 式两边同时乘以 $1'$ 可得出：

$$\rho_t = \rho_{t-1}(\alpha_{t-1}\mathrm{B}_t 1') \tag{5.7}$$

从上式，我们能得出：

$$\ln L_T = \sum_{t=1}^{T}\ln(\rho_t/\rho_{t-1}) = \sum_{t-1}^{T}\ln(\alpha_{t-1}\mathrm{B}_t 1') \tag{5.8}$$

Γ 和 $\mathrm{P}(x_t)$ 是 m 阶矩阵，$\mathrm{B}_t = \Gamma\mathrm{P}(x_t)$ 。上述对数似然函数 $\ln L_T$ 的计算可以直接纳入到下面的递归算法中。初始值 $f_0 = \delta$ ，$h_0 = 0$ 。对于 $t = 1,2,\cdots T$ ，循环计算以下步骤：$v_t = \alpha_{t-1}\Gamma\mathrm{P}(x_t)$ ；$u_t = v_t 1'$ ；$h_t = h_{t-1} + \ln u_t$ ；$\alpha_t = u_t/v_t$ 。这样最后求得的 h_T 即为 $\ln L_T$ 。其中，h_t 是累计对数似然函数的标量，相当于 ρ_t/ρ_{t-1} 。v 和 α_t 是 m 维向量，u 是标量。这个过程能在很多情况下避免下溢问题。

在泊松－隐蔽马尔科夫模型中，转移矩阵 Γ 和泊松分布参数向量 λ 的元素是有非负约束的。比如，Γ 的每一行元素之和都应等于 1。因此，当我们对似然函数进行估计时，就需要解决有约束条件的最优问题，而不是无约束条件的最优问题。这个问题我们已在上一章进行过讨论。

HMM 的似然函数是一个关于多个参数的复杂方程，常常含有多个局部极值。我们的目标是找到似然函数关于参数总体的最值，但是并没有一个判断最大化算法是否已经达到总体最优的方法。由于算法依赖于初始值的设定，因此所得到的局部最优值很有可能不是总体最优值，后面将要介绍的 EM 算法也会出现这种问题。因此，可以考虑使用多个初始值，并观察在不同情况下出现的最优值是否相同。

对 HMM 的某些参数来说，找到比较恰当的初始值是不难的。例如，估计两状态泊松 - 隐蔽马尔科夫模型时，若样本方差为5，则可以尝试4 和6 或者3 和7 作为两个状态均值的初始值。此外，基于分位数的估计策略也是可行的。例如，如果模型有三种状态，可以将样本的上分位数、中位数、下分位数作为三个状态均值的初始值。

第二节　EM 算法

一、EM 算法介绍

对 HMM 似然函数进行估计比较有效的方法是 EM 算法，用这种算法时我们需要用到向前概率和向后概率，后面章节中解码和状态预测的内容中也会用到这两种概率。EM 算法也被称为 Baum - Welch 算法，用于齐次马尔科夫链的 HMM 的参数估计，并不要求马尔科夫链一定是稳态的，因此并没有假设 $\delta\Gamma = \delta$ 。因此，除了给定状态下的分布所含参数 λ 和转移矩阵 Γ，这种算法也要估计初始分布 w 。

根据式 5. 1 我们定义行向量 f_t 如下：

$$f_t = w\mathrm{P}(x_1)\Gamma\mathrm{P}(x_2)\cdots\Gamma\mathrm{P}(x_t) = w\mathrm{P}(x_1)\prod_{s=2}^{t}\Gamma\mathrm{P}(x_s) \tag{5.9}$$

其中 $t = 1,2,\cdots,T$, w 表示马尔科夫链的初始分布。我们已经给出前向概率 f_t 中元素，但并没有对此给出详细说明。本节内容将表明 f_t 的第 j 个元素 $f_t(j)$ 确实是一个概率函数，而且联合概率是 $P(X_1 = x_1,X_2 = x_2,\cdots,X_t = x_t,S_t = j)$ 。

我们也需要用到后向概率 b_t ：

$$b'_t = \Gamma\mathrm{P}(x_{t+1})\Gamma\mathrm{P}(x_{t+2})\cdots\Gamma\mathrm{P}(x_T)1' = (\prod_{s=t+1}^{T}\Gamma\mathrm{P}(x_s))1' \tag{5.10}$$

$t = T$ 时定义 $b_T = 1$ 。实际上，$b(j)$ 的第 j 个元素 $b_t(j)$ 也是一个条件概率，表示 $P(X_{t+1} = x_{t+1},X_{t+2} = x_{t+2},\cdots,X_T = x_T \mid S_t = j)$ 。基于以上分析，我们可以得到：

$$f_t(j)b_t(j) = P(X_1^T = x_1^T, S_t = j) \tag{5.11}$$

根据前面公式 5.1 已经给出的定义 $f_{t+1} = f_t \Gamma P(x_{t+1})$，$f_{t+1}(j)$ 具体表达式为：

$$\begin{aligned} f_{t+1}(j) &= (\sum_{i=1}^{m} f_t(i) r_{ij}) p_j(x_{t+1}) \\ &= \sum_{i=1}^{m} f_t(i) r_{ij} p_j(x_{t+1}) \\ &= \sum_{i=1}^{m} P(X_1^t, S_t = i) \cdot P(S_{t+1} = j \mid S_t = i) \cdot P(X_{t+1} = x_t \mid S_{t+1} = j) \\ &= \sum_{i=1}^{m} P(X_1^t, S_t = i, S_{t+1} = j) \\ &= P(X_1^t, S_{t+1} = j) \end{aligned}$$

对于 $t = 1,2,\cdots,T$ 和 $j = 1,2,\cdots,m$，我们可以得到下面的结论：

$$f_t(j) = P(X_1^t = x_1^t, S_t = j) \tag{5.12}$$

同样道理，根据之前也已经给出 b_t 的定义可知：

$$b_t = \Gamma \mathrm{P}(x_{t+1}) b_{t+1}$$

可以同样得到下面的结论：

$$b_t(i) = P(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \cdots, X_T = x_T \mid S_t = i)$$

$$b_t(i) = P(X_{t+1}^T = x_{t+1}^T \mid S_t = i) \tag{5.13}$$

我们将 EM 算法运用到 HMM 中，得到：

$$f_t(j)b_t(j) = P(X_1^T = x_1^T, S_t = i)$$

因此，对每个 t，有 $\alpha_t \beta_t = P(X_1^T = x_1^T) = L_T$。这是因为：

$$L_T = \delta \mathrm{P}(x_1) \Gamma \mathrm{P}(x_2) \cdots \Gamma \mathrm{P}(x_t) \Gamma \mathrm{P}(x_{t+1}) \Gamma \mathrm{P}(x_{t+2}) \cdots \Gamma \mathrm{P}(x_T) 1' = \alpha_t \beta'_t$$

基于以上对 L_T 的分析，我们可以得到如下结论：

$$P(S_t = j \mid X_1^T = x_1^T) = \alpha_t(j) \beta_t(j) / L_T \tag{5.14}$$

$$P(S_{t-1} = j, S_t = k \mid X_1^T = x_1^T) = \alpha_{t-1}(j) r_{ij} p_k(x_t) \beta_t(k) / L_T \tag{5.15}$$

由 HMM 中状态变量虽然服从马尔科夫过程，但无法被观察到，所以在 HMM 参数估计中，人们常常会把这些状态当作缺失数据，然后采用 EM 算法来找到参数的

极大似然估计。事实上，开创性的工作是由复兴科技的 Leonard Baum 等人在 20 世纪 70 年代初期完成的，Dempster（1977）等人在其基础上展开了进一步的研究。

当一些数据丢失时，EM 算法能够用迭代方法来进行最大似然估计，也就是说，EM 方法不是简单地使可观察值的似然函数最大化，而是使弥补了不可观测数据后的完整数据的对数似然函数（CDLL）最大化。完整数据的对数似然函数是在观察值和缺失数据基础上，并且含有参数 θ 的对数似然函数。

二、EM 算法具体步骤

1. 选择参数 θ 的初始值。

2. 基于观察值和当前估计值 θ，计算缺失数据的条件期望值，即计算出 CDLL 中缺失数据的条件期望。

3. 把 CDLL 中的缺失数据用条件期望替代，进行极大似然估计。

4. 重复步骤 2、3 的操作直到一些收敛性判断成立，比如直到 θ 收敛为止。

此时的 θ 就是似然函数的最优解。在某些情况下，这个最优解可能只是一个局部最大值或鞍点。

EM 算法的关键理念不仅仅是强调 CDLL 中缺失的数据本身由它们的条件期望值取代，更为重要的是含有这些缺失数据条件期望的 CDLL 函数。

在 HMM 中，人们经常用 0－1 变量来描述马尔科夫状态序列 s_1，s_2，…，s_T：

$u_j(t) = 1$，当且仅当 $S_t = j$。

$v_{jk}(t) = 1$，当且仅当 $S_{t-1} = j$ 并且 $S_t = k$。

HMM 的完整数据对数似然函数 CDLL，也就是由观察值 x_1，x_2，…，x_T 和缺失数据 s_1，s_2，…，s_T 组成的似然函数，可以写成：

$$\ln(P(x_1^T, s_1^T)) = \ln\left(w_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t} \prod_{t=1}^{T} p_{s_t}(x_t)\right)$$

$$= \ln w_{s_1} + \sum_{t=2}^{T} n\gamma_{s_{t-1}, s_t} + \sum_{t=1}^{T} \ln p_{s_t}(x_t) \tag{5.16}$$

所以：

$$\ln(P(x_1^T, s_1^T))$$

$$= \sum_{j=1}^{m} \hat{u}_j(1)\ln w_j + \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\sum_{t=2}^{T}\hat{v}_{jk}(t)\right)\ln\gamma_{jk} + \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\ln p_j(x_t)$$

$$= \mathrm{I1} + \mathrm{I2} + \mathrm{I3}$$

其中 γ_{jk} 表示转换概率矩阵 Γ 的第 (j,k) 个元素。w 是马尔科夫链的初始分布，即 S_1 的分布，但却未必是稳态分布。如果只从一个观察值来估计初始分布是不合理的，况且对于马尔科夫链来说其状态本身又是不可观测的。EM 算法可以解决这一问题，假设马尔科夫链不仅是齐次的而且是稳态的，这样就有 $w = \delta$，初始值的估计问题就迎刃而解。

三、HMM 的 EM 算法具体过程

E 步骤：用基于观察值 x_1^T 所得到的条件期望 $\hat{v}_{jk}(t)$ 和 $\hat{u}_j(t)$ 替代所有的 $v_{jk}(t)$ 和 $u_j(t)$，得到完整数据对数似然函数 CDLL，即：

$$\hat{u}_j(t) = P(S_t = j \mid x^{(T)}) = f_t(j)b_t(j)/L_T \qquad (5.17)$$

$$\hat{v}_{jk}(t) = P(S_{t-1} = j, S_t = k \mid x_1^T) = f_{t-1}(j)\gamma_{jk}p_k(x_t)b_t(j)/L_T \qquad (5.18)$$

注意到我们需要前向概率和后向概率来计算 $\hat{v}_{jk}(t)$ 和 $\hat{u}_j(t)$，但这里没有假设潜在变量 S_t 是具有马尔科夫链稳态性的。

M 步骤：用 $\hat{v}_{jk}(t)$ 和 $\hat{u}_j(t)$ 替代 $v_{jk}(t)$ 和 $u_j(t)$ 得到的完整数据的对数似然函数 CDLL，对其进行最优化，该最优化包括如下三组参数：初始分布 w，转移概率矩阵 Γ 和状态依赖型分布的参数（如简单泊松－隐蔽马尔科夫模型的 $\lambda_1, \cdots, \lambda_m$）。

CDLL 最大化过程可以分成三个独立的部分，因为第一项（I1）只依赖于初始分布 w，第二项（I2）是状态转移概率矩阵 Γ，第三项（I3）是状态依赖型分布的参数。我们可以采取以下三个独立步骤来求 CDLL 最大化：

1. $\sum_{j=1}^{m} \hat{u}_j(1)\ln w_j$ 关于 w 求最大化；

2. $\sum_{j=1}^{m}\sum_{k=1}^{m}\left(\sum_{t=2}^{T}\hat{v}_{jk}(t)\right)\ln\gamma_{jk}$ 关于 Γ 求最大化；

3. $\sum_{j=1}^{m} \sum_{t=1}^{T} \hat{u}_j(t) \ln p_j(x_t)$ 关于分布参数求最大化。

以上三个部分最优化后的结果如下：

1. $w_j = \frac{\hat{u}_j(1)}{\sum_{j=1}^{m} \hat{u}_j(1)} = \hat{u}_j(1)$

证明：

$$\max \sum_{j=1}^{m} \hat{u}_j(1) \ln\delta_j$$

$$s.t.\ \ w_1 + w_2 + \cdots + w_m = 1$$

以上极值问题的拉格朗日函数为：

$$L = \sum_{j=1}^{m} \hat{u}_j(1) \log w_j + \lambda(1 - w_1 - w_2 - \cdots - w_m)$$

$$\frac{\partial L}{\partial w_j} = \frac{\hat{u}_j(1)}{w_j} - \lambda = 0$$

$$w_j = \frac{\hat{u}_j(1)}{\lambda}$$

$$\sum_{j=1}^{m} w_j = \frac{\sum_{j=1}^{m} \hat{u}_j(1)}{\lambda} = 1$$

$$\lambda = \sum_{j=1}^{m} \hat{u}_j(1)$$

$$w_j = \hat{u}_j(1) / \sum_{j=1}^{m} \hat{u}_j(1) = \hat{u}_j(1)$$

2. $\gamma_{jk} = \alpha_{jk} / \sum_{k=1}^{m} \alpha_{jk}$ 其中 $\alpha_{jk} = \sum_{t=2}^{T} \hat{v}_{jk}(t)$ ，这个问题我们已在前面专门讨论过。

3. 第三步最大化可能容易也可能困难，这由状态依赖型分布的性质决定。这个问题实际上是一个和分布有关的极大似然估计问题。对于泊松分布和正态分布而言，解析解是可以得到的。但对其他的一些分布而言，比如伽马分布和负二项分布，数值解也是可以通过 M 步骤来得到。

值得注意的是，前向和后向概率的计算出现下溢或上溢问题，EM 算法采取加权的办法能够防止这个问题的发生，或者至少减少这样的风险。下面我们着重讨论

完整数据对数似然函数 CDLL 第三项（I3）的最大化问题，这一部分的 CDLL 为：

$$\sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\log p_j(x_t)$$

四、泊松分布

泊松分布的概率分布 $p_j(x) = e^{-\lambda_j}\lambda_j^{x}/x!$，所以，对以下对数似然函数进行最优化：

$$\begin{aligned} L &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\ln p_j(x_t) \\ &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\ln e^{-\lambda_j}\lambda_j^{x_t}/x_t! \\ &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)(-\lambda_j + x_t\ln\lambda_j - \ln x_t!) \end{aligned}$$

求导得：

$$\frac{\partial L}{\partial \lambda_j} = \sum_{t=1}^{T}\hat{u}_j(t)(-1 + x_t/\lambda_j) = 0$$

$$\hat{\lambda}_j = \sum_{t=1}^{T}\hat{u}_j(t)x_t / \sum_{t=1}^{T}\hat{u}_j(t) \tag{5.19}$$

五、正态分布

正态分布的概率密度 $p_j(x) = (2\pi\sigma_j^2)^{-1/2}\exp\left[-\frac{1}{2\sigma_j^2}(x-\mu_j)\right]^2$，对以下对数似然函数进行最优化：

$$\begin{aligned} L &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\ln p_j(x_t) \\ &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\ln\left[(2\pi\sigma_j^2)^{-1/2}\exp\left(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2\right)\right] \\ &= \sum_{j=1}^{m}\sum_{t=1}^{T}\hat{u}_j(t)\left[-\frac{1}{2}\ln(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}(x-\mu_j)^2\right] \end{aligned}$$

求 μ_j 和 σ_j^2 求导，可得：

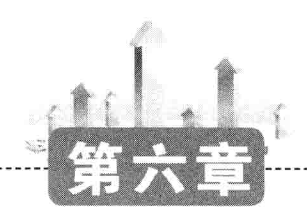
$$\frac{\partial L}{\partial \mu_j} = \sum_{t=1}^{T} \hat{u}_j(t) \times \frac{x_t - \mu_j}{\sigma_j^2} = 0$$

$$\frac{\partial L}{\partial \sigma_j^2} = \sum_{t=1}^{T} \hat{u}_j(t) \times \frac{(x_t - \mu_j)^2 - \sigma_j^2}{\sigma_j^4} = 0$$

综上可得：

$$\hat{\mu}_j = \sum_{t=1}^{T} \hat{u}_j(t) x_t / \sum_{t=1}^{T} \hat{u}_j(t) \tag{5.20}$$

$$\hat{\sigma}_j^2 = \sum_{t=1}^{T} \hat{u}_j(t) \ (x_t - \mu_j)^2 / \sum_{t=1}^{T} \hat{u}_j(t) \tag{5.21}$$



第六章

隐蔽马尔科夫模型应用与模型选择

本章讨论以下几个十分重要的问题：HMM 观测值的条件分布 $P(X_t = x \mid X^{(-t)} = x^{(-t)}) = \sum_i w_i(t) p_i(x)$，即给定除 t 时刻之外所有观测值，求 t 时刻观测值的条件分布；HMM 的预测分布 $P(X_{T+h} = x \mid X_1^T = x_1^T) = \sum_i \xi_i(t) p_i(x)$，即给定所有观测值求 h 期之后观测值的分布；HMM 解码问题，包括局部解码 $P(S_t = i \mid X_1^T = x_1^T) = f_t(i) b_t(i) / L_T$ 和全局解码 $P(S_1^T = s_1^T \mid X_1^T = x_1^T)$，即给定观测值条件下求某时刻 t 或整个时间范围内状态变量的分布；HMM 状态预测问题 $P(S_{T+h} = i \mid X_1^T = x_1^T) = f_T \Gamma^h(\cdot, i) / L_T$，即给定观测值条件下 h 期之后状态变量的分布。

注意，本章没有假设马尔科夫链 $\{S_t\}$ 是稳态的，而是仅做了齐次性假定。用行向量 w 表示初始分布，即 S_1 的分布，这里并未假定它是稳态分布。当然，本章的结论对于稳态隐蔽马尔科夫模型这一特殊情况也是成立的，这时 δ 既是初始分布又是稳态分布。

第一节　条件分布

现在，我们推导在给定所有其他观测值下，X_t 的条件分布：$P(X_t = x \mid X^{(-t)} = x^{(-t)})$。这里，我们把除了 t 时刻之外其他所有时刻的观测值记为 $X^{(-t)}$，即：

$$X^{(-t)} \equiv (X_1, \cdots, X_{t-1}, X_{t+1}, \cdots, X_T)$$

使用前面讨论过的关于前向概率、后向概率的定义以及 HMM 的似然函数，我们可以很快得到下式：

对于 $t = 2,3,\cdots,T$，有：

$$P(X_t = x \mid X^{(-t)} = x^{(-t)}) = \frac{w\mathrm{P}(x_1)\mathrm{B}_2\cdots\mathrm{B}_{t-1}\Gamma\mathrm{P}(x)\mathrm{B}_{t+1}\cdots\mathrm{B}_T1'}{w\mathrm{P}(x_1)\mathrm{B}_2\cdots\mathrm{B}_{t-1}\ \Gamma\mathrm{B}_{t+1}\cdots\mathrm{B}_T1'}$$

$$\propto w\mathrm{P}(x_1)\mathrm{B}_2\cdots\mathrm{B}_{t-1}\Gamma P(x)\mathrm{B}_{t+1}\cdots\mathrm{B}_T1'$$

$$\propto f_{t-1}\Gamma\mathrm{P}(x)b'_t \tag{6.1}$$

由上一章内容可知，$f_t = w\mathrm{P}(x_1)\mathrm{B}_2\cdots\mathrm{B}_t$，$b'_t = \mathrm{B}_{t+1}\cdots\mathrm{B}_T1'$ 且 $b_T = 1$，$\mathrm{B}_t = \Gamma\mathrm{P}(x_t)$。

当 $t=1$ 时，$P(X_1 = x \mid X^{(-1)} = x^{(-1)}) = \dfrac{w\mathrm{P}(x)\mathrm{B}_2\cdots\mathrm{B}_T1'}{w\,\mathrm{I}\mathrm{B}_2\cdots\mathrm{B}_T1'} \propto w\mathrm{P}(x)b'_1$ (6.2)

以上条件分布是 HMM 的两个似然值之比：分子是观测值 x_t 被替换为 x 后的似然值，分母是 x_t 缺失时的似然值。上面两式中的条件概率都具备如下形式：行向量 $f_{t-1}\Gamma$ 乘以一个 $m\times m$ 对角阵 $\mathrm{P}(x) = diag(p_1(x),\cdots,p_m(x))$，再乘以一个列向量 b'_t。于是，对于 $t = 1,2,\cdots,T$，我们有：

$$P(X_1 = x \mid X^{(-t)} = x^{(-t)}) \propto \sum_{i=1}^{m} d_i(t)p_i(x) \tag{6.3}$$

其中，$d_i(t)$ 是向量 $f_{t-1}\Gamma$ 第 i 个元素与向量 b_t 的第 i 个元素之积；因此，

$$P(X_t = x \mid X^{(-t)} = x^{(-t)}) \propto \sum_{i=1}^{m} \tau_i(t)p_i(x) \tag{6.4}$$

其中，混合权重 $\tau_i(t) = d_i(t)/\sum_{j=1}^{m} d_j(t)$。

第二节 预测分布

现在我们来讨论 HMM 的预测分布问题。具体来讲，我们将推导在给定 $X_1^T = x_1^T$ 的条件下，X_{T+h} 条件分布的两种表达。其中，h 被称为预测范围。我们还是重点讨论离散的情形；连续情形与离散情形基本相同，只是把概率函数替换为密度函数。

对于离散观测值的 HMM，其预测分布 $P(X_{T+h} = x \mid X_1^T = x_1^T)$ 与之前讨论的条件分布 $P(X_t = x \mid X^{(-t)} = x^{(-t)})$ 很相似，并且计算方法本质上也是相同的——即为两个似然值之比：

$$P(X_{T+h} = x \mid X_1^T = x_1^T) = \frac{P(X_1^T = x_1^T, X_{T+h} = x)}{P(X_1^T = x_1^T)}$$

$$= \frac{w\mathrm{P}(x_1)\mathrm{B}_2\mathrm{B}_3\cdots\mathrm{B}_T\Gamma^h\mathrm{P}(x)1'}{w\mathrm{P}(x_1)\mathrm{B}_2\mathrm{B}_3\cdots\mathrm{B}_T1'}$$

$$= \frac{f_T\Gamma^h\mathrm{P}(x)1'}{f_T1'} \tag{6.5}$$

由上一章可知 $\alpha_T = f_T/f_T1'$，则有：

$$P(X_{T+h} = x \mid X_1^T = x_1^T) = \alpha_T\Gamma^h\mathrm{P}(x)1' \tag{6.6}$$

因此，预测分布可以写成状态依赖概率分布的混合：

$$P(X_{T+h} = x \mid X_1^T = x_1^T) = \sum_{i=1}^{m} \varpi_i(h)p_i(x) \tag{6.7}$$

其中，权重 $\varpi_i(h)$ 是向量 $\alpha_T\Gamma^h$ 的第 i 个元素。

第三节　解码

在语音识别研究中，研究序列发生时所处的状态是颇受关注的，并将其称为解码问题。在 HMM 中，解码是指在已知观测值的条件下推断未知状态变量的过程，分为局部解码（Local Decoding）和全局解码（Global Decoding）。更为具体地讲，t 时刻状态的局部解码是指在某个时刻每个状态最可能发生的概率 $P(S_t = i \mid X_1^T = x_1^T) = f_t(i)b_t(i)/L_T$，而全局解码是指某一状态序列最可能发生的概率 $P(S_1^T = s_1^T \mid X_1^T = x_1^T)$。下面将分别阐述这两种解码。

为推导马尔可夫链在 t 时刻最可能的状态，我们要用到如下结论：

$$f_t(i)b_t(i) = P(X_1^T = x_1^T, S_t = i) \tag{6.8}$$

因此，给定可观察值 X_1^T 的情况下，S_t 的条件分布为：

$$P(S_t = i \mid X_1^T = x_1^T) = \frac{P(S_t = i, X_1^T = x_1^T)}{P(X_1^T = x_1^T)} = \frac{f_t(i)b_t(i)}{L_T} \qquad (i = 1,2,\cdots,m) \tag{6.9}$$

其中，L_T 可以通过加权前向概率的方法来计算。此方法对防止计算乘积 $f_t(i)b_t(i)$ 的数值下溢也是必要的。

对于每个时间 $t \in \{1,\cdots,T\}$，给定观察值 X_1^T，最可能发生的状态 i_t^* 被定义为：

$$i_t^* = \underset{i=1,\cdots,m}{\operatorname{argmax}} P(S_t = i \mid X_1^T = x_1^T) \tag{6.10}$$

这种方法分别对于每个时间点 t，都使条件概率 $P(S_t = i \mid X_1^T = x_1^T)$ 最大化，从而确定了最为可能的状态，所以被称为**局部解码**。

在很多应用中，比如语音识别，相对于局部解码所得到的每一个时刻 t 的最可能状态，人们可能对最有可能发生的状态序列更感兴趣。这里提醒大家一下，这个

状态序列是不可观测的，或者说是隐蔽的，所以需要以概率的方式来进行推测。全局解码不是分别对每个时刻 t 来最大化 $P(S_t = i \mid X_1^T = x_1^T)$，而是寻找状态序列 $s_1, s_2, \cdots, s_T$，以最大化条件概率：

$$P(S_1^T = s_1^T \mid X_1^T = x_1^T)$$

或者等价地，使下面的联合概率最大化：

$$P(S_1^T = s_1^T, X_1^T = x_1^T) = w_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1},s_t} \prod_{t=1}^{T} p_{s_t}(x_t)$$

这是一个与局部解码不同的最大化问题，被称为**全局解码**。局部和全局解码的结果通常十分相似，但并不完全相同。

我们可以采用数值计算的方法最大化上式来求得最优状态序列 $s_1, s_2, \cdots, s_T$。问题是这需要对 m^T 个函数进行估计，除非 T 特别小，否则这种方法显然是不可行的。在解决这一问题方面比较有效的算法是 Viterbi 算法（1967），可以用来确定最优状态序列。下面简单介绍 Viterbi 算法的具体步骤。

首先定义：

$$\xi_{1i} = P(S_1 = i, X_1 = x) = w_i p_i(x_1) \tag{6.11}$$

$$\xi_{ti} = \max_{s_1, s_2, \cdots, s_{t-1}} P(S_1^{t-1} = s_1^{t-1}, S_t = i, X_1^T = x_1^T) \quad t = 2, 3, \cdots, T \tag{6.12}$$

显然，概率 ξ_{tj} 满足以下递归：

$$\xi_{tj} = [\max_i(\xi_{t-1,i}\gamma_{ij})] p_j(x_t) \quad t = 2, 3, \cdots, T; i = 1, 2, \cdots, m \tag{6.13}$$

由于该式的计算对 T 来说是线性的，所以存在一种计算 ξ_{tj} 值的有效方法。最优状态序列 $i_1, i_2, \cdots, i_T$ 可以从下式递归得到：

$$i_T = \underset{i=1,2,\cdots,m}{\operatorname{argmax}} \xi_{Ti} \tag{6.14}$$

$$i_t = \underset{i=1,2,\cdots,m}{\operatorname{argmax}} (\xi_{ti}\gamma_{i,t+1}) \quad t = T-1, T-2, \cdots, 1 \tag{6.15}$$

由于全局解码中所用到的极大似然函数是概率的乘积，所以可以选择对其取自然对数以避免数值下溢。Viterbi 算法将极大似然函数写成概率的对数形式，因此可以避免数值下溢。此外，Viterbi 算法在极大似然函数中还使用了加权权重的办法，这与我们前面计算 $\tau_i(t)$ 方法很相似。这种加权缩放法也可用于这里，只

需将矩阵 $\{\xi_{ti}\}$ 的各行进行缩放，使得每行的各元素之和为 1。Viterbi 算法既可运用于稳态马尔科夫链也可用于非稳态马尔科夫链，因而无须假设初始分布 w 是稳态分布。

第四节　状态预测

在前面的讨论中，我们推导出了给定观测值 X_1^T 下状态 S_t 的条件分布，就时间点来讲，只考虑了现在或过去的状态。下面将给出 $t>T$ 时未来状态 S_t 的条件分布，即进行状态预测。

给定观察值 $x_1, x_2, \cdots, x_T$，可以得出下面一系列有关未来、现在和过去状态的表述：

$$P(S_t = i \mid X_1^T = x_1^T) = \begin{cases} f_t \Gamma^{t-T}(\cdot, i)/L_T & t > T \text{ 时} \quad \text{状态预测} \\ f_T(i)/L_T & t = T \text{ 时} \quad \text{平滑过程} \\ f_t(i) b_t(i)/L_T & 1 \le t < T \text{ 时} \quad \text{过滤过程} \end{cases}$$

其中，$\Gamma^{t-T}(\cdot, i)$ 表示矩阵 Γ^{t-T} 的第 i 列。过滤和平滑部分和前面所描述的状态概率是相同的，事实上由于对任何 i 都有 $b_T(i)=1$，这两部分可以合而为一。状态预测部分仅仅是对 $t>T$ 即未来的一种概括，即：

$$P(S_{T+h} = i \mid X_1^T = x_1^T) = \frac{f_T \Gamma^h(\cdot, i)}{L_T} = \alpha_T \Gamma^h(\cdot, i) \qquad i = 1, 2, \cdots, m \quad (6.16)$$

其中 $\alpha_T = f_T / f_T 1'$。需要注意，当 $h \to \infty$时，$\alpha_T \Gamma^h$ 趋近于马尔可夫链稳态分布。

第五节　模型选择标准

在有 m 个状态的隐蔽马尔科夫模型中，增加状态的个数一般会改善模型的拟合程度。但是，这也会带来待估参数数量以平方的速度增加，所以在拟合效果和参数数量之间存在一个权衡。因此，需要确定一个选择模型的标准。

在某些情况下，对状态依赖型分布或者转移概率矩阵做出假设以减少参数的个数是一种明智的选择。

下面，我们对 HMM 的模型选择标准做一个简单介绍。

在使用 HMM 时，我们会经常遇到如下这些问题：比如如何选择状态 m 的个数，或者如何选择状态依赖型分布的参数，比如在泊松分布和负二项分布之间如何抉择。这需要我们为模型选择确定一些标准。

我们至少有两种方法进行模型选择。第一种是使用统计的方法，将模型选择标准简化为 AIC（Akaike Information Criterion）：

$$\mathrm{AIC} = -2\ln L + 2K \tag{6.17}$$

式中 $\ln L$ 为拟合模型的对数最大似然函数，K 为模型中参数的个数。第一项的拟合值随着状态 m 的个数增加而减小，第二项是一个惩罚项，随着状态 m 的个数增加而增大。

第二种是运用贝叶斯的方法进行模型选择。在这种情形下，我们使用与 AIC 不同的惩罚项，这种方法形成的模型选择标准被称为 BIC（Bayesian Information Criterion）：

$$\mathrm{BIC} = -2\ln L + K \cdot \ln T \tag{6.18}$$

上式中 $\ln L$ 和 K 的含义和 AIC 中相同，T 是观察值的个数。与 AIC 相比，当 $T > e^2$ 时，BIC 的惩罚项有更大的权重，因此，很多人喜欢使用 BIC 的方法来进行模型选择。当然，与 AIC 相比，BIC 一般比较适合参数较少的模型选择问题。

第三部分

马尔科夫状态转换模型

在讨论马尔科夫状态转换模型（Markov – Switching Model，MS – AR 模型）之前，先让我们认识一下非稳定时间序列模型。这里的稳定性指的是模型参数在不同时间段之间的相对固定性。在对时间序列的回归分析中，如果任取一个时间段，将这个序列向前或向后移动，其回归模型参数仍保持不变，则此时的回归模型是稳定的；如果在回归过程中，某一时期内的参数和另一时期内的参数明显不同，则称此时的回归模型是非稳定的。这里重点考察的非稳定时间序列模型是马尔科夫状态转换模型，即对整体数据的描述不存在稳定的时间序列模型，但是对局部数据的描述存在稳定的时间序列模型，且随着局部数据的推移往往需要在几个模型间不断切换，即时间序列中存在转换（Switch）。下面将对非稳定模型展开研究。

目前已经有大量文献研究如何判定在某时间点前后是否存在不同的模型参数组来描述数据生成过程。普林斯顿大学邹至庄教授提出了邹氏检验法，用于判断模型在预先给定的时点是否发生了变化。这种方法的特点在于把时间序列数据分成两部分，并检验模型是否在其分界点已发生结构性变化。在此基础上，利用 F 检验来检验由前一部分 n 个数据求得的参数与由后一部分 m 个数据求得的参数是否相等，由此判断模型是否发生了变化。

但问题是，在实际操作中，研究者通常不清楚时间序列结构变化的具体时间点，因此，需要推断这些转折点（Turning Point）发生的时间。早期研究都只考虑仅有一个未知转折点的时间序列数据问题，例如 Quandt（1958），Farley 和 Hinich（1970）。后来，Goldfeld 和 Quandt（1973）开始考虑允许多个未知转折点的时间序列数据问题。此外，这些模型也逐步取消了转换概率外生性的假设，即决定转换点发生的因素包括在模型内部。Goldfeld 和 Quandt（1973）的模型创新性地引入马尔科夫过程，明确假设状态转换服从马尔科夫过程。特别是，Hamilton（1989）的马

尔科夫状态转换模型（State - dependent Markov - Switching Model）引起了人们对模型中状态变量的关注。Hamilton（1989）模型可以看作是 Goldfeld 和 Quandt（1973）模型在有关状态依赖的自回归模型方面的进一步拓展。有关这个问题的深入探讨，请查阅 Hamilton（1993）所著的《Time Series Analysis》一书。

对于马尔科夫状态转换模型而言，复杂之处在于随机变量 y_t 是一个状态依赖变量。具体来说是指 y_t 的分布或分布的参数取决于状态变量 S_t 。我们只能得到随机变量 y_t 基于状态变量 S_t 的条件分布 $f(y_t \mid S_t)$ 。对于状态变量 S_t 来说，可能是可观测的，也可能是不可观测的；可能是相互独立的，也可能是相互影响的。对于相互影响情况下的 S_t ，我们常常假设其服从马尔科夫过程。同样，y_t 本身可以是相互独立的，也可以是序列相关的。所以，尽管我们对状态变量施加了马尔科夫性要求，但是模型仍然具有高度的灵活性和适用性。

本书的第三部分将分以下四种情况来详细讨论马尔科夫状态转换模型：

		状态变量 S_t	
		相互独立	马尔科夫
序列变量 y_t	无自相关	第七章第一节（状态不可观测）	第七章第二节（状态不可观测）
	自相关		第八章第一节（状态可观测） 第八章第二节（状态不可观测）

第九章将基于第八章第二节的马尔科夫状态转换（MS - AR）模型讨论参数的估计问题。



第七章

序列不相关数据的马尔科夫状态转换模型

序列不相关意味着过去观测值对未来观测值的预测没有任何价值，用数学语言表达即 $y_t \neq f(y_{t-1}, \cdots, y_1, y_0)$ 。在没有状态转换时，我们的回归模型是：

$$y_t = x_t\beta + \varepsilon_t \qquad \varepsilon_t \sim i.i.d.N(0,\sigma^2)$$

其中 x_t 是关于 $1 \times k$ 维外生变量的向量。我们可以用极大似然估计法估计模型的系数：

$$\ln L = \sum_{1}^{T} \ln(f(y_t))$$

通过对 β 和 σ^2 求一阶导并令其等于零，可以得到对数似然值的最大化值，并且求解参数的估计值。具体过程如下：

由于 $\varepsilon_t \sim i.i.d.N(0,\sigma^2)$ ，可知：

$$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\varepsilon_t^2}{2\sigma^2})$$

因为 x_t 是已知的观测值，β 也是一个固定值，所以 y_t 密度函数可以写为：

$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y_t - x_t\beta)^2}{2\sigma^2})$$

有关这个问题的求解，我们已在第一章中进行过详细讨论。接下来，我们考虑在此基础上加入状态转换后的模型变化情况：

$$y_t = x_t\beta_{s_t} + \varepsilon_t \qquad t = 1,2,\cdots,T \tag{7.1}$$

$$\varepsilon_t \sim N(0,\sigma_{s_t}^2) \qquad S_t = 0 \text{ 或 } 1 \tag{7.2}$$

$$\beta_{s_t} = \beta_0(1 - s_t) + \beta_1 s_t \tag{7.3}$$

$$\sigma_{s_t}^2 = \sigma_0^2(1 - s_t) + \sigma_1^2 s_t \tag{7.4}$$

这样，模型参数出现了一些变化，不再是固定的取值，而是在两组取值间反复转换。通过上述模型设置可知，在 $S_t = 0$ 时，$\beta_{s_t} = \beta_0$，$\sigma_{s_t}^2 = \sigma_0^2$；而在 $S_t = 1$ 时，$\beta_{s_t} = \beta_1$，$\sigma_{s_t}^2 = \sigma_1^2$。如果 $\beta_0 = \beta_1$ 并且 $\sigma_0^2 = \sigma_1^2$，那么这意味着稳态时间序列模型可以看作是马尔科夫状态转换模型的特例；如果两组参数不相等，那么就意味着模型参数在两组取值之间不断转换。基于以上模型设置，我们需要对以下两种情况分别进行讨论：

第一种情况指状态变量 S_t 直接可观测。这是指在 t 时可以直接观测得到 S_t 的值，从而使问题大大简化。此时，我们可以使用极大似然估计方法来分别估计，即在 $S_t = 0$ 时，估计 β_0 和 σ_0^2；在 $S_t = 1$ 时，估计 β_1 和 σ_1^2。计算步骤如下：

1. 给定 $y_T, y_{T-1}, \cdots, y_1$ 以及 $S_T = s_T, S_{T-1} = s_{T-1}, \cdots, S_1 = s_1$，可以得到观测值关于状态变量的条件分布

$$f(y_T \mid S_T = s_T), f(y_{T-1} \mid S_{T-1} = s_{T-1}), \cdots, f(y_1 \mid S_1 = s_1) \tag{7.5}$$

2. 条件密度函数为 $f(y_t \mid S_t = s_t) = \dfrac{1}{\sqrt{2\pi\sigma_{s_t}^2}}\exp\left(-\dfrac{(y_t - x_t\beta_{s_t})^2}{2\sigma_{s_t}^2}\right)$； (7.6)

3. 似然函数为 $L = f(y_T \mid S_T = s_T) \cdot f(y_{T-1} \mid S_{T-1} = s_{T-1}) \cdot \cdots \cdot f(y_1 \mid S_1 = s_1)$ (7.7)

4. 对数似然函数为 $\ln L = \sum_1^T \ln(f(y_t \mid S_t = s_t))$ (7.8)

5. 通过分别对 $\beta_0, \beta_1, \sigma_0^2, \sigma_1^2$ 求一阶导，最大化上式

$$\max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2} \ln L = \max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2} \sum_1^T \ln(f(y_t \mid S_t = s_t))$$

第二种情况是指状态变量 S_t 不可观测。在这种情况下，部分参数的取值会影响到状态的识别，从而影响到观测值的分布，并连同其他参数共同影响极大似然值的大小。根据所能获取的历史信息 $y_1^{t-1} = (y_1, y_2, ..., y_{t-1})$，我们要估计 $f(S_t \mid y_1^{t-1})$。

计算过程如下：

1. 由贝叶斯定理可得$f(y_t,S_t)=f(y_t\mid S_t)\cdot f(S_t)$，在加入条件 y_1^{t-1} 后

$$f(y_t,S_t\mid y_1^{t-1})=f(y_t\mid S_t,y_1^{t-1})\cdot f(S_t\mid y_1^{t-1}) \tag{7.9}$$

2. 由于 S_t 是估计得到的，而 y_1^{t-1} 是已知的，所以在计算极大似然值时是基于历史观测值来推测条件概率，用$f(y_t\mid y_1^{t-1})$来计算极大似然值，而不用第一种情况中所使用的$f(y_t\mid S_t=s_t)$。用类似积分消除的方法将$f(y_t\mid y_1^{t-1})$写成$f(y_t,S_t\mid y_1^{t-1})$项对 S_t 加总的形式，并以此达到计算 $f(y_t\mid y_1^{t-1})$ 的目的。因为 S_t 只能等于 0 或 1，所以：

$$\begin{aligned}
&f(y_t\mid y_1^{t-1})=\sum_{s_t=0}^{1}f(y_t,S_t\mid y_1^{t-1})\\
&=f(y_t,S_t=0\mid y_1^{t-1})+f(y_t,S_t=1\mid y_1^{t-1})\\
&=f(y_t\mid S_t=0,y_1^{t-1})\cdot f(S_t=0\mid y_1^{t-1})+f(y_t\mid S_t=1,y_1^{t-1})\cdot f(S_t=1\mid y_1^{t-1})\\
&=\sum_{s_t=0}^{1}f(y_t\mid S_t,y_1^{t-1})\cdot f(S_t\mid y_1^{t-1})\\
&=\sum_{s_t=0}^{1}f(y_t\mid S_t)\cdot f(S_t\mid y_1^{t-1})\\
&=\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp(-\frac{(y_t-x_t\beta_0)^2}{2\sigma_0^2})P(S_t=0\mid y_1^{t-1})\\
&\quad+\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-\frac{(y_t-x_t\beta_1)^2}{2\sigma_1^2})P(S_t=1\mid y_1^{t-1})
\end{aligned} \tag{7.10}$$

3. 似然方程为

$$\begin{aligned}
L&=\prod_{t=1}^{T}f(y_t\mid y_1^{t-1})\\
&=\prod_{t=1}^{T}\sum_{s_t=0}^{1}f(y_t\mid S_t,y_1^{t-1})\cdot P(S_t\mid y_1^{t-1})
\end{aligned} \tag{7.11}$$

两边取自然对数，可得：

$$\ln L=\sum_{t=1}^{T}\ln(\sum_{s_t=0}^{1}f(y_t\mid S_t,y_1^{t-1})\cdot P(S_t\mid y_1^{t-1})) \tag{7.12}$$

4. 通过对 $\beta_0,\beta_1,\sigma_0^2,\sigma_1^2$ 求一阶导，最大化上式

$$\max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2} \ln L = \max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2} \sum_{t=1}^{T} \ln\left(\sum_{s_t=0}^{1} f(y_t \mid S_t, y_1^{t-1}) \cdot P(S_t \mid y_1^{t-1})\right)$$

在状态变量 S_t 不可观测的情况下，问题显得更为复杂。下面假设 S_t 不可观测，但是在给定 S_t 的条件下 y_t 不存在序列相关性，并在此假设基础上针对 S_t 之间相互独立和满足马尔科夫性两种情况展开对模型估计的讨论。

第一节　序列不相关且状态相互独立的转换模型

有关序列 y_t 不相关并且状态 S_t 不可观测且相互独立的转换模型，具体假设如下：

1. 假设 y_t 满足序列不相关，即有 $y_t \neq f(y_{t-1}, \cdots, y_1, y_0)$；

2. 假设每一期的状态变量 S_t 是无法观测的未知变量，因此需要对 S_t 的变化做出进一步的假设；

3. 假设 S_t 每期之间的变化是互相独立的，即 S_t 的取值与 S_1^{t-1} 都不相关。

我们要研究的问题是，在已知数据集 $\{y_1^{t-1}\}$ 的前提下，来估计模型参数以及不可观测的状态序列 S_1^{t-1}，并以此来推测 S_t 和 y_t 等。

与以往模型不同的是，这里需要假设 S_t 的值是未知的，我们不能从数据中直接观测得到，因此需要对 S_t 的值做出推断。比如，我们可以根据之前的数据和信息 $\{y_1^{t-1}\}$ 来估计 S_t，即 $f(S_t \mid y_1^{t-1})$。具体过程如下：

1. 假设 S_t 的条件概率为 $f(S_t = 0 \mid y_1^{t-1}) = p$，且有 $f(S_t = 1 \mid y_1^{t-1}) = 1 - p$。这样就可以得到在状态变量独立假设下 S_t 的取值概率；

2. 对于 y_t 的条件分布，我们仍然假设 $f(y_t \mid S_t) = \dfrac{1}{\sqrt{2\pi\sigma_{s_t}^2}} e^{-\frac{(y_t - x_t\beta)^2}{2\sigma_{s_t}^2}}$；

3. 基于以上假设，我们就可以得出似然函数。在状态独立假设中，S_t 是我们根据之前期的信息推断出来的，不能简单地直接用 S_t 的信息来确定 y_t，而要用 $\{y_1^{t-1}\}$

来确定 y_t：

$$L = f(y_t \mid y_1^{t-1}) \cdot f(y_{t-1} \mid y_1^{t-2}) \cdot \cdots \cdot f(y_1 \mid y_0) \tag{7.13}$$

其中，$f(y_t \mid y_1^{t-1}) = \sum_{S_t=0}^{1} f(y_t, S_t \mid y_1^{t-1})$。

因为 S_t 是观测值所处事件空间的测度，令 $\Omega(\cdot)$ 表示变量所对应的事件集合，则有：

$$\Omega(S_t = 0) \cap \Omega(S_t = 1) = \varnothing \text{ 且 } \Omega(S_t = 0) \cup \Omega(S_t = 1) = \Omega$$

所以，我们可以得出：

$$\begin{aligned}(y_t \cap \Omega(S_t = 0)) \cup (y_t \cap \Omega(S_t = 1)) &= y_t \cap (\Omega(S_t = 0) \cup \Omega(S_t = 1)) \\ &= y_t \cap \Omega = y_t\end{aligned}$$

这样，就可以得到：

$$f(y_t \mid y_1^{t-1}) = \sum_{S_t=0}^{1} f(y_t \mid S_t, y_1^{t-1}) \cdot f(S_t \mid y_1^{t-1}) \tag{7.14}$$

其中加总项中的第一部分表示给定历史观测值和当期状态变量之后当期观测值的似然值，第二部分表示给定历史观测值之后当期状态变量取值的概率。

我们已经完全利用了 y_1^{t-1} 中的信息来推断 S_t，因此：

$$f(y_t \mid y_1^{t-1}) = \sum_{S_t=0}^{1} f(y_t \mid S_t) \cdot f(S_t \mid y_1^{t-1}) \tag{7.15}$$

有关 y_t 的似然函数为：

$$\begin{aligned}L &= f(y_t \mid y_1^{t-1}) \cdot f(y_{t-1} \mid y_1^{t-2}) \cdot \cdots \cdot f(y_1 \mid y_0) \\ &= \prod_{t=1}^{T} \sum_{S_t=0}^{1} f(y_t \mid S_t) \cdot f(S_t \mid y_1^{t-1}) \\ &= \prod_{t=1}^{T} \sum_{S_t=0}^{1} \left[\frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} e^{-\frac{(y_t - x_t\beta_{S_t})^2}{2\sigma_{S_t}^2}} \cdot f(S_t \mid y_1^{t-1}) \right]\end{aligned} \tag{7.16}$$

两边取自然对数，可得：

$$\ln L = \ln \prod_{t=1}^{T} \sum_{S_t=0}^{1} \left[\frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} e^{-\frac{(y_t - x_t\beta_{st})^2}{2\sigma_{S_t}^2}} \cdot f(S_t \mid y_1^{t-1}) \right]$$

$$= \sum_{t=1}^{T} \ln\left\{ \sum_{S_t=0}^{1} \left[\frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} e^{-\frac{(y_t - x_t\beta_{st})^2}{2\sigma_{S_t}^2}} \cdot f(S_t \mid y_1^{t-1}) \right] \right\}$$

$$= \sum_{t=1}^{T} \ln\left[\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \cdot p + \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \cdot (1-p) \right] \tag{7.17}$$

我们上面的做法是，先依靠 y_1^{t-1} 的信息来估计 $S_t = 0$ 和 $S_t = 1$ 的概率。根据不同的状态概率，我们可以确定每个状态下 y_t 的似然分布，再将两种状态相加。由最大似然估计的原理，我们可以得到以下似然函数：

$$\max \ln L = \max \ln \sum_{t=1}^{T} \left[\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \cdot p + \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \cdot (1-p) \right]$$

在求解以上方程时，由于概率 $p \in (0,1)$，所以应该对 p 加以约束，但这样的话将会大大增加求解的难度。因此，我们对 P 作如下处理：

$$P(S_t = 0) = p = \frac{\exp(\tilde{p})}{1 + \exp(\tilde{p})} \tag{7.18}$$

则有：

$$P(S_t = 1) = 1 - p = 1 - \frac{\exp(\tilde{p})}{1 + \exp(\tilde{p})} = \frac{1}{1 + \exp(\tilde{p})} \tag{7.19}$$

这里，$\tilde{p}$ 是一个无约束的参数。如果 S_t 随机估计值不依赖于任何其他的外生变量，我们就可以得到 $P(S_t = j \mid y_1^{t-1}) = P(S_t = j)$。那么，通过求解以上最大似然函数，我们可以求出 $\beta_0, \beta_1, \sigma_0^2, \sigma_1^2, \tilde{p}$ 几个未知参数。

接下来，我们再考虑一个更为复杂的情况，$P(S_t)$ 与前期信息 y_1^{t-1} 不相关，但可能与其他外生或者前定变量相关。假定有 n 个外生或者前定变量，我们将这些变量放到 $(t-1) \times n$ 阶矩阵 Z_{t-1} 中，$\tilde{p}$ 对 Z_{t-1} 回归可以得到：

$$\tilde{p} = \alpha_p + Z'_{t-1} \cdot \beta_p \tag{7.20}$$

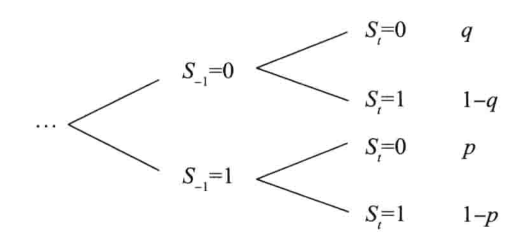
$$P(S_t = 0) = p = \frac{\exp(\alpha_p + Z'_{t-1} \cdot \beta_p)}{1 + \exp(\alpha_p + Z'_{t-1} \cdot \beta_p)} \tag{7.21}$$

$$P(S_t = 1) = 1 - p = \frac{1}{1 + \exp(\alpha_p + Z'_{t-1} \cdot \beta_p)} \tag{7.22}$$

这样，似然函数中就包含 $\beta_0,\beta_1,\sigma_0^2,\sigma_1^2,\alpha_p,\beta_p$ 六个未知参数。通过求解极大似然函数，我们就可以得到这些参数的估计。

第二节　序列不相关马尔科夫状态转换模型

之前所讨论的两种转换情况中，S_t 都是根据 y_1^{t-1} 直接推断出的；然而在实际情况中，S_t 往往难以被观察到，但可以根据 S_{t-1} 推测出 S_t 。我们常常假设状态 S_t 的变化服从马尔科夫过程，这就是马可科夫状态转换模型。下面我们开始研究这个模型，仍然假设只有两个状态，即 $S_t = 1$ 或者 $S_t = 0$ 。



我们假设状态变量 S_t 有如下的转换关系：

$$P(S_t = 1 | S_{t-1} = 1) = p$$

$$P(S_t = 0 | S_{t-1} = 0) = q$$

处理马尔科夫状态转换模型的方法与上一节所讨论的解决办法相似，共需要如下四个步骤：

1. 初始化过程：给定 $w_0 = P(S_0 = 0 | y_0)$ 和 $w_1 = P(S_1 = 0 | y_0)$ 。y_0 表示到时间 0 为止得到的所有信息。

2. 预测过程：对于 $i, j = 0$ 或 1 ，在给定 $P(S_{t-1} = i | y_1^{t-1})$ 的条件下求

$$\begin{aligned} P(S_t = j | y_1^{t-1}) &= \sum_{i=0}^{1} P(S_t = j, S_{t-1} = i | y_1^{t-1}) \\ &= \sum_{i=0}^{1} P(S_t = j | S_{t-1} = i) P(S_{t-1} = i | y_1^{t-1}) \end{aligned} \tag{7.23}$$

3. 更新过程：

在第 t 期，y_t 变为已知，$y_1^t = (y_1^{t-1}, y_t)$。这就需要更新对 $P(S_t = j \mid y_1^{t-1})$ 的估计，或在第 t 次迭代后，其概率变为：

$$P(S_t = j \mid y_1^t) = P(S_t = j \mid y_1^{t-1}, y_t)$$

由于 $P(S_t = j \mid y_t) = \dfrac{P(S_t = j, y_t)}{P(y_t)}$，等式两边同时加入条件 y_1^{t-1} 得：

$$\begin{aligned} P(S_t = j \mid y_1^{t-1}, y_t) &= \frac{P(S_t = j, y_t \mid y_1^{t-1})}{P(y_t \mid y_1^{t-1})} \\ &= \frac{P(y_t \mid S_t = j, y_1^{t-1}) P(S_t = j \mid y_1^{t-1})}{\sum_{j=0}^{1} P(y_t, S_t = j \mid y_1^{t-1})} \\ &= \frac{P(y_t \mid S_t = j, y_1^{t-1}) P(S_t = j \mid y_1^{t-1})}{\sum_{j=0}^{1} P(y_t \mid S_t = j, y_1^{t-1}) P(S_t = j \mid y_1^{t-1})} \end{aligned} \tag{7.24}$$

4. 当 $t + 1 > T$ 时，计算结束。否则，重复步骤1－3，由 $P(S_t = j \mid y_1^t)$，可得 $P(S_{t+1} = j \mid y_1^t)$。

重复以上四个步骤，即可得到 $P(S_t = j \mid y_1^{t-1})$ $(t = 1, 2, \cdots, T)$。

在上述计算过程中的 $t=1$ 时，需要对初始值 $P(S_0 \mid y_0)$ 予以赋值。可用以下步骤求得该初始值。

1. 对转移概率进行如下假设：

$$P(S_t = 0 \mid S_{t-1} = 0) = p = \frac{\exp(\tilde{p})}{1 + \exp(\tilde{p})}$$

$$P(S_t = 1 \mid S_{t-1} = 1) = q = \frac{\exp(\tilde{q})}{1 + \exp(\tilde{q})}$$

$$P(S_t = 1 \mid S_{t-1} = 0) = 1 - p$$

$$P(S_t = 0 \mid S_{t-1} = 1) = 1 - q$$

2. 我们再假设 $P(S_0 = 0) = w_0$，$P(S_0 = 1) = 1 - w_0$

$P(S_{t+1} = 0) = P(S_{t+1} = 0 \cap S_t = 1) + P(S_{t+1} = 0 \cap S_t = 0)$

$$= P(S_{t+1}=0 \mid S_t=1)\cdot P(S_t=1)+P(S_{t+1}=0 \mid S_t=0)\cdot P(S_t=0)$$

$$= (1-q)\cdot(1-P(S_t=0))+p\cdot P(S_t=0)$$

$$= (1-q)+(p+q-1)\cdot P(S_t=0)$$

3. 利用以上方程进行类推，可得：

$$P(S_1=0)=(1-q)+(p+q-1)\cdot P(S_0=0)$$

$$=(1-q)+(p+q-1)\cdot w_0$$

$$P(S_2=0)=(1-q)+(p+q-1)\cdot P(S_1=0)$$

$$=(1-q)+(1-q)(p+q-1)+(p+q-1)^2\cdot w_0$$

4. 递推可得：

$$P(S_n=0)=(1-q)\sum_{i=0}^{n-1}(p+q-1)^i+(p+q-1)^n\cdot w_0$$

$$=\frac{1-q-(1-q)(p+q-1)^n}{2-p-q}+(p+q-1)^n\cdot w_0$$

$$=\frac{1-q}{2-p-q}+(p+q-1)^n\left(w_0-\frac{1-q}{2-p-q}\right)$$

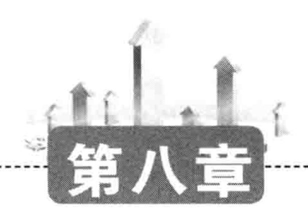
由于 $|p+q-1|<1$，得：

$$\lim_{n\to\infty}P(S_n=0)=\frac{1-q}{2-p-q}$$

因此，应当假设初始值 $P(S_0=0 \mid y_0)$ 为：

$$P(S_0=0 \mid y_0)=\frac{1-q}{2-p-q}$$

同理：$P(S_0=1 \mid y_0)=\dfrac{1-p}{2-p-q}$。这样求出的概率被称为无条件概率或稳态概率。这时，似然函数就包含 β_0，β_1，σ_0，σ_1，p，q 六个未知参数。通过求解极大似然函数，我们就可以得到这些系数的估计值。



第八章 序列自相关的马尔科夫状态转换模型

第一节　序列自相关且状态可观测的马尔科夫状态转换模型

我们下面考虑的情况是，y_t 序列自相关，同时假设 S_t 为可观测变量。此时，$y_t = f(y_{t-1}, y_{t-2}, \cdots, y_1)$ 。

一般地，观测值序列自相关且含有 M 个状态的马尔科夫转换模型可以写为：

$$\varphi(L)(y_t - \mu_{s_t}) = \varepsilon_t \qquad \varepsilon_t \sim i.i.d.N(0, \sigma_{s_t}^2)$$

$$P(S_{t+1} = j \mid S_t = i) = p_{ij} \qquad i, j = 1, 2, \cdots, M$$

$$\mu_{S_t} = \mu_1 S_{1t} + \mu_2 S_{2t} + \cdots + \mu_M S_{Mt}$$

$$\sigma_{S_t}^2 = \sigma_1^2 S_{1t} + \sigma_2^2 S_{2t} + \cdots + \sigma_M^2 S_{Mt}$$

后面两个式子分别表示均值的转换和方差的转换。其中 $\varphi(L)$ 为滞后算子。在某一时点 t，参数处于且仅处于一个状态，其取值取决于状态变量 S_t 。当 $S_t = m$ 时，$S_{mt} = 1$ ，否则 $S_{mt} = 0$ 。

为简化起见，我们从 y_t 为 AR（1）的情况入手。此时，

$$(y_t - \mu_{s_t}) = \varphi_1(y_{t-1} - \mu_{s_{t-1}}) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_{s_t}^2)$$

由于 S_t 是可观测的，因而以上完全可以看作是一个虚拟变量模型。假设 y_t 的密度函数为：

$$f(y_t \mid y_1^{t-1}, S_t, S_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} \exp\left(-\frac{\{(y_t - \mu_{S_t}) - \varphi_1(y_{t-1} - \mu_{S_{t-1}})\}^2}{2\sigma_{S_t}^2}\right)$$

有关 y_t 的似然函数为：

$$L = f(y_t \mid y_1^{t-1}, S_t, S_{t-1}) \cdot f(y_{t-1} \mid y_1^{t-2}, S_{t-1}, S_{t-2}) \cdot \cdots \cdot f(y_1 \mid y_0, S_1, S_0)$$

两边同时取自然对数，可得对数形式的最大似然函数为：

$$\ln L = \sum_{t=1}^{T} \ln f(y_t \mid y_1^{t-1}, S_t, S_{t-1})$$

由于状态 S_t 的可观测性，我们很容易利用以上似然函数求出参数的最优估计值。

第二节　序列自相关和状态不可观测的马尔科夫状态转换模型

在上一节例子中，y_t 为 $AR(1)$ 过程，且在给定过去 $t-1$ 期的信息 y_1^{t-1} 的情况下，如果要描述 y_t 的密度函数，还需要变量 S_t 与 S_{t-1} 的信息。当在 t 期状态变量 S_t 与 S_{t-1} 不可观测的时候，问题就变得较为复杂。为了解决这样的问题，我们运用与上一节相似的方法，但是不再考虑 y_t 和 S_t 的联合密度函数，而考虑使用 y_t，S_t 和 S_{t-1} 的联合密度函数。模型设置如下：

给定 $y_t, y_{t-1}, \cdots, y_1$ 且数据产生过程服从 $AR(1)$ 模型，即：

$$(y_t - \mu_{S_t}) = \varphi_1(y_{t-1} - \mu_{S_{t-1}}) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_{S_t}^2) \tag{8.1}$$

其中 $S_t, S_{t-1}, \cdots, S_1$ 不可观测，但服从一阶马尔科夫过程；

$$f(y_t \mid y_1^{t-1}, S_t, S_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} \exp\left(-\frac{\{(y_t - \mu_{S_t}) - \varphi_1(y_{t-1} - \mu_{S_{t-1}})\}^2}{2\sigma_{S_t}^2}\right) \tag{8.2}$$

在这样的情况下，分两个步骤来建立极大似然函数并对参数进行估计。具体步骤如下：

1. 在以过去信息 φ_{t-1} 为条件的情况下，得到 y_t，S_t 和 S_{t-1} 的联合密度函数：

$$f(y_t,S_t,S_{t-1} \mid y_1^{t-1}) = f(y_t \mid y_1^{t-1},S_t,S_{t-1})P(S_t,S_{t-1} \mid y_1^{t-1}) \tag{8.3}$$

其中，$f(y_t \mid y_1^{t-1},S_t,S_{t-1})$ 可由式 8.2 得到。

2. 为了利用上式求得 $f(y_t \mid \varphi_{t-1})$，需要将包含 S_t 和 S_{t-1} 所有可能值的联合密度函数进行求和，即：

$$\begin{aligned} f(y_t \mid y_1^{t-1}) &= \sum_{S_t=1}^{M} \sum_{S_{t-1}=1}^{M} f(y_t,S_t,S_{t-1} \mid y_1^{t-1}) \\ &= \sum_{S_t=1}^{M} \sum_{S_{t-1}=1}^{M} f(y_t \mid S_t,S_{t-1},y_1^{t-1})P(S_t,S_{t-1} \mid y_1^{t-1}) \end{aligned} \tag{8.4}$$

在上式中，边际密度 $f(y_t \mid y_1^{t-1})$ 是 M^2 个条件密度的加权平均，权重分别是 $P(S_t,S_{t-1} \mid y_1^{t-1})$，其中 $i,j = 1,2,\cdots,M$。

3. 综上可得似然函数：

$$\ln L = \sum_{t=1}^{T} \ln\{\sum_{S_t=1}^{M} \sum_{S_{t-1}=1}^{M} f(y_t \mid S_t,S_{t-1},y_1^{t-1})P(S_t,S_{t-1} \mid y_1^{t-1})\} \tag{8.5}$$

在计算以上对数似然函数方程最大值并求解参数之前，我们仍然需要解决如何计算 $P(S_t,S_{t-1} \mid y_1^{t-1})$（$t = 1,2,\cdots,T$）的问题。

第三节　滤波过程

$P(S_t,S_{t-1} \mid y_1^{t-1})$（$t = 1,2,\cdots,T$）可按照下面两个步骤循环得到，我们将这个过程称为 Hamilton 滤波（Filtering）。

第一步被称为预测过程（Forecasting），即以 y_1^{t-1} 信息为条件，在给定 $P(S_{t-1} = i \mid y_1^{t-1})$ $i = 1,2,\cdots,M$ 的条件下，求 $P(S_t,S_{t-1} \mid y_1^{t-1})$。计算的方法如下：

$$P(S_t,S_{t-1} \mid y_1^{t-1}) = P(S_t = j \mid S_{t-1} = i)P(S_{t-1} = i \mid y_1^{t-1}) \tag{8.6}$$

其中，对于 $i,j = 1,2,\cdots,M$ 有 $P(S_t = j \mid S_{t-1} = i)$，表示马尔科夫过程的转移概率。

第二步被称为更新过程（Updating），即在给定 $P(S_t,S_{t-1} \mid y_1^{t-1})$ 和新增观测值 y_t

的条件下求 $P(S_t,S_{t-1}\mid y_1^t)$ 。在 t 期结束后，我们就可通过观察得到 y_t ，信息集由 y_1^{t-1} 更新为 y_1^t 。所以就可以将所要计算的概率按照以下方法更新：

$$
\begin{aligned}
P(S_t,S_{t-1}\mid y_1^t) &= P(S_t,S_{t-1}\mid y_1^{t-1},y_t) \\
&= \frac{f(S_t=j,S_{t-1}=i,y_t\mid y_1^{t-1})}{f(y_t\mid y_1^{t-1})} \\
&= \frac{f(y_t\mid S_t=j,S_{t-1}=i,y_1^{t-1})P(S_t=j,S_{t-1}=i\mid y_1^{t-1})}{\sum_{i=1}^{M}\sum_{j=1}^{M}f(y_t\mid S_t=j,S_{t-1}=i,y_1^{t-1})P(S_t=j,S_{t-1}=i\mid y_1^{t-1})}
\end{aligned}
\tag{8.7}
$$

同时得到：

$$
P(S_t=j\mid y_1^{t-1}) = \sum_{i=1}^{M}P(S_t=j,S_{t-1}=i\mid y_1^{t-1}) \tag{8.8}
$$

以备下一个循环中第一步预测过程所用。

重复以上两个步骤，可以使我们计算出 $P(S_t,S_{t-1}\mid y_1^{t-1})$ 。从 $t=1$ 开始过滤，我们可以使用稳态概率或无条件概率作为初始值。在两个状态的情况下，一阶马尔科夫转换的稳态概率如下：

$$
w_1 = P(S_0=1\mid y_0) = \frac{1-p_{22}}{2-p_{22}-p_{11}} \tag{8.9}
$$

$$
w_2 = P(S_0=2\mid y_0) = \frac{1-p_{11}}{2-p_{22}-p_{11}} \tag{8.10}
$$

对 Hamilton 滤波的总结，可以参考下面计算过程：

1. 计算初始值 w_1 和 w_2 ，对 $t=1,2,\cdots,T$ 循环计算下面步骤 2 和 4，并根据式 8.3 求得各个时期的对数似然值；

2. 根据式 8.6 计算预测过程；

3. 根据式 8.3 计算更新前的对数似然值；

4. 获取新的观测值数据信息后，根据式 8.7、式 8.8 计算更新过程。

完成以上步骤，求得对数似然函数方程式 8.5，该对数似然方程是关于未知参数的方程，可以通过极大似然估计等方法求得参数的估计值。我们始终要明确的是，

马尔科夫状态转换模型的推断包括以下部分：一是通过极大似然估计法估计模型的系数；二是推断状态序列 S_t ，$t = 1,2,\cdots,T$ 。

上面对 y_t 服从 $AR(1)$ 过程的状态转换模型的分析可以很容易地推广到状态转换的一般 $AR(k)$ 模型。在这种情况中，由于 $S_t,S_{t-1},\cdots S_{t-k}$ 在模型中是不可观测的，我们考虑 $y_t,S_t,S_{t-1},\cdots S_{t-k}$ 的联合密度函数。由此，y_t 的边际密度函数 $f(y_t \mid y_1^{t-1})$ 为 M^{k+1} 个条件概率密度的加权平均。

在一般情况中，对状态变量 S_t 在不同的信息集下进行推断，我们分别可以得到滤波概率和平滑概率。滤波概率是指根据从期初到 t 期的信息 y_1^t 对推断 S_t 的 $P(S_t \mid y_1^{t-1})$ 。平滑概率是指根据全部样本信息 y_1^T 对 S_t 做出推断 $P(S_t \mid y_1^T)$ ，这可以通过下一节的平滑过程得到。

第四节　平滑过程

在给出模型参数估计的情况下，我们可以根据样本的全部信息来推断 S_t 的值。前文给出的是滤波概率，即为 $P(S_t = i \mid y_1^t)$ ，$t = 1,2,\cdots,T$ 。而这里我们将要讨论的是平滑（Smoothing）概率，即 $P(S_t = j \mid y_1^T)$ ，$t = 1,2,\cdots,T$ 。现将平滑过程和过滤过程进行简单对比。

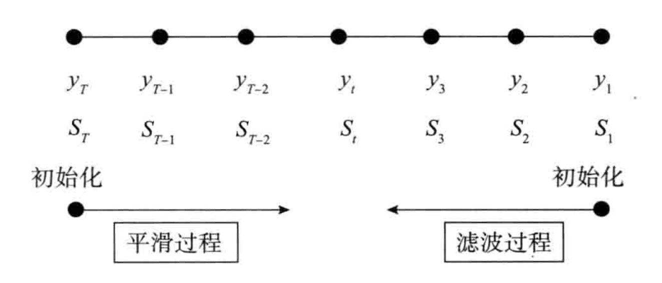


图 8－1　平滑过程与过滤过程对比图

表 8－1 平滑过程与过滤过程对比表

	计算方向	概率	适用情况
平滑过程	从第 T 期逆时间顺序计算	$P(S_t = i \mid y_1^t)\ t = 1,2,\cdots,T$ 根据全部样本信息计算概率	基于全部数据
过滤过程	从第 1 期顺时间顺序计算	$P(S_t = j \mid y_1^T)\ t = 1,2,\cdots,T$ 根据过去样本信息计算概率	基于实时数据

下面，就平滑概率的 Kim（1998）算法进行讨论。假设所分析模型仍为一阶自回归的马尔科夫状态转换模型，以下是基于全部信息 ψ_T 的 $S_t = j$ 和$S_{t+1} = k$ 联合概率的计算过程。

$$
\begin{aligned}
& P(S_t = j \mid S_{t+1} = k, y_1^T) \\
&= P(S_{t+1} = k \mid y_1^T) \times P(S_t = j \mid S_{t+1} = k, y_1^T) \\
&= P(S_{t+1} = k \mid y_1^T) \times P(S_t = j \mid S_{t+1} = k, y_1^t) \\
&= \frac{P(S_{t+1} = k \mid y_1^T) \times P(S_t = j \mid y_1^t) \times P(S_{t+1} = k \mid S_t = j)}{P(S_{t+1} = k \mid y_1^t)}
\end{aligned}
\tag{8.11}
$$

$$
P(S_t = j \mid y_1^T) = \sum_{k=1}^{M} P(S_t = j, S_{t+1} = k \mid y_1^T) \tag{8.12}
$$

式8.11 中，分子 $P(S_t = j \mid y_1^t)$ 和分母中 $P(S_{t+1} = k \mid S_t = j)$ 均可由 Hamilton 滤波过程求得。给出最后一期 $P(S_T \mid y_1^T)$，则通过式 8.11 和式 8.12 对 $t = 1,2,\cdots,T$ 进行迭代计算出其他各期 $P(S_t \mid y_1^T)$ $(t = 1,2,\cdots,T-1)$。

现对式 8.11 的第二行和第三行的等价关系进行推导，即需要证明：

$$
P(S_t = j \mid S_{t+1} = k, y_1^T) = P(S_t = j \mid S_{t+1} = k, y_1^t)
$$

对 $T > t$，定义 y_{t+1}^T 为从 $t+1$ 期到 T 期的观测值向量。因此有：

$$
\begin{aligned}
& P(S_t = j \mid S_{t+1} = k, y_1^T) \\
&= P(S_t = j \mid S_{t+1} = k, y_{t+1}^T, y_1^t) \\
&= \frac{f(S_t = j, y_{t+1}^T \mid S_{t+1} = k, y_1^t)}{f(y_{t+1}^T \mid S_{t+1} = k, y_1^t)}
\end{aligned}
$$

$$= \frac{P(S_t = j \mid S_{t+1} = k, y_1^t) f(y_{t+1}^T \mid S_{t+1} = k, S_t = j, y_1^t)}{f(y_{t+1}^T \mid S_{t+1} = k, y_1^t)}$$

$$= P(S_t = j \mid S_{t+1} = k, y_t) \tag{8.13}$$

式 8.13 成立的条件是$f(y_{t+1}^T \mid S_{t+1} = k, S_t = j, y_1^t) = f(y_{t+1}^T \mid S_{t+1} = k, y_1^t)$，这意味着如果 S_{t+1} 已知，那么对 y_{t+1} 分布的预测不再需要 S_t 的信息，因为 S_{t+1} 和 y_1^t 已经包含了 S_t 的信息。在平滑算法中上式成立，所以有式 8.11 和式 8.12。

上述平滑算法的推导可以一般化为一个含有马尔科夫转换的 k 阶自回归模型，该模型由 Hamilton（1989）提出。

$$P(S_{t-k+1}, \cdots, S_t, S_{t+1} \mid y_1^T)$$

$$= P(S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T) \times P(S_{t-k+1} \mid S_{t-k+2}, \cdots, S_t, S_{t+1}, y_1^T)$$

$$= \frac{P(S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T) \times P(S_{t-k+1}, S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T)}{P(S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T)}$$

$$= \frac{P(S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T) \times P(S_{t-k+1}, S_{t-k+2}, \cdots, S_t \mid y_1^T) \times P(S_{t+1} \mid S_t)}{P(S_{t-k+2}, \cdots, S_t, S_{t+1} \mid y_1^T)} \tag{8.14}$$

这个含有马尔科夫转换的 k 阶自回归模型看起来很复杂，但是不难证明对 $t \leq T - k + 1$ 式 8.14 可以分解成式 8.1。引人注意的是，Kim（1998）算法要比 Hamilton（1989）的算法和 Lam（1990）的算法简单得多，也在很大程度上节约了计算时间。

第五节　马尔科夫转换模型中 S_t 状态的持续期

转移概率矩阵 $\begin{bmatrix} P_{11} & \cdots & P_{M1} \\ \vdots & \ddots & \vdots \\ P_{1M} & \cdots & P_{MM} \end{bmatrix}$ 中的对角线上的元素包含着有关持续期的重要信息。相关的问题是：如果我们知道当期 S_t 处于 j 值的状态（$S_t = j$），那么这种状态平均会持续多久？

定义 D 为 S_t 处于 j 状态的持续期数，则有：

如果 $S_t = j$，$S_{t+1} \neq j$，则持续 1 期，$D = 1$，$P(D = 1) = (1 - p_{jj})$；

如果 $S_t = S_{t+1} = j$，$S_{t+2} \neq j$，则持续 2 期，$D = 2$，$P(D = 2) = p_{jj}(1 - p_{jj})$；

如果 $S_t = S_{t+1} = S_{t+2} = j$，$S_{t+3} \neq j$，则持续 3 期，$D = 3$，$P(D = 3) = p_{jj}{}^2(1 - p_{jj})$；

如果 $S_t = S_{t+1} = S_{t+2} = S_{t+3} = j$，$S_{t+4} \neq j$，则持续 4 期，$D = 4$，$P(D = 4) = p_{jj}{}^3(1 - p_{jj})$。

以此类推，则持续值为 j 期的期望为：

$$
\begin{aligned}
E(D) &= \sum_{j=1}^{\infty} P(D = j) \\
&= 1 \times P(S_{t+1} \neq j \mid S_t = j) \\
&\quad + 2 \times P(S_{t+1} = j, S_{t+2} \neq j \mid S_t = j) \\
&\quad + 3 \times P(S_{t+1} = j, S_{t+2} = j, S_{t+3} \neq j \mid S_t = j) \\
&\quad + 4 \times P(S_{t+1} = j, S_{t+2} = j, S_{t+3} = j, S_{t+4} \neq j \mid S_t = j) \\
&\quad + \cdots \\
&= 1 \times (1 - p_{jj}) + 2 \times p_{jj}(1 - p_{jj}) + 3 \times p_{jj}^2(1 - p_{jj}) + \cdots \\
&\quad + (n - 1) \times p_{jj}^{n-2}(1 - p_{jj}) + n \times p_{jj}^{n-1}(1 - p_{jj}) \\
&= 1 - p_{jj} + 2p_{jj} - 2p_{jj}^2 + 3p_{jj}^2 - 3p_{jj}^3 + \cdots + (n - 1)p_{jj}^{n-2} \\
&\quad - (n - 1)p_{jj}^{n-1} + np_{jj}^{n-2} - np_{jj}^n \\
&= (1 + p_{jj} + p_{jj}^2 + \cdots + p_{jj}^{n-1}) - np_{jj}^n \\
&= \frac{1 - p_{jj}^n}{1 - p_{jj}} - np_{jj}^n
\end{aligned}
$$

当 $n \to \infty$ 时 $p_{jj}^n \to 0$，所以有：

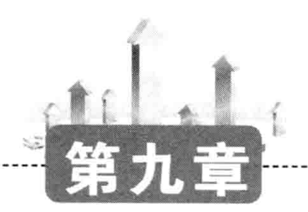
$$
\mathrm{E}(D) = \sum_{j=1}^{\infty} P(D = j) = \frac{1}{1 - p_{jj}} \tag{8.15}
$$

例如，在 Hamilton（1989）对美国季度 GNP 变量的研究模型中，一共含有两种状态，状态 1 表示经济衰退，状态 2 表示经济繁荣，而 GNP 增长率满足两个状态转

换的 MS－AR（4）模型，对转换概率 P_{11}和 P_{22}的估计值是 0.7750 和 0.9049。根据式 8.15，对这两种状态的持续期估计分别是：

$$\frac{1}{1-0.7750}=4.08\ ,\frac{1}{1-0.9040}=10.42$$

也就是说，平均意义上来看，一个衰退期和繁荣期分别持续 4.08 和 10.42 个季度。



第九章

MS－AR 模型的估计方法

第一节　MS－AR模型参数估计初步

相对于普通的线性模型来说，MS－AR 模型的估计更为复杂。我们首先通过一个简单的例子来说明估计 MS－AR 模型的复杂性。假设 MS－AR 模型中观测值 y_t 不具有序列相关性，即 $y_t \neq f(y_{t-1},\cdots,y_1)$，$S_t$ 服从独立转换过程且不能被观测到，在这一简化 MS－AR 模型中对数似然函数的最大化问题为：

$$\max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2,p} \ln L = \max_{\beta_0,\beta_1,\sigma_0^2,\sigma_1^2,p} \ln \sum_{t=1}^{T} \left[\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \cdot p + \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \cdot (1-p) \right] \tag{9.1}$$

对数似然函数对各个参数求一阶微分，从而可以估计求得模型中的参数，具体步骤和解法如下。

一、β_0 的最优解

式 9.1 对 β_0 求一阶微分得：

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{\partial \left[\sum_{t=1}^{T} \ln f(y_t \mid \varphi_{t-1}) \right]}{\partial \beta_0} = \sum_{t=1}^{T} \frac{\partial \ln f(y_t \mid \varphi_{t-1})}{\partial \beta_0}$$

$$= \sum_{t=1}^{T} \left\{ \frac{p}{f(y_t \mid \varphi_{t-1})} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \left[-\frac{1}{2\sigma_0^2}(y_t - x_t\beta_0) \times 2(-x_t) \right] \right\}$$

$$\Rightarrow \sum_{t=1}^{T} \frac{e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}}}{f(y_t \mid \varphi_{t-1})}(y_t - x_t\beta_0)x_t = 0$$

如令 $z_t = \frac{e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_0^2}}}{f(y_t \mid \varphi_{t-1})}$，则：

$$\sum_{t=1}^{T} z_t(y_t - x_t\beta_0)x_t = 0$$

$$\Rightarrow \sum_{t=1}^{T} z_t y_t x_t = \sum_{t=1}^{T} z_t x_t^2 \beta_0$$

$$\Rightarrow \hat{\beta}_0 = \frac{\sum_{t=1}^{T} z_t y_t x_t}{\sum_{t=1}^{T} z_t x_t^2} = \frac{\sum_{t=1}^{T} (\sqrt{z_t} x_t)(\sqrt{z_t} y_t)}{\sum_{t=1}^{T} (\sqrt{z_t} x_t)^2}$$

由 $y_t = x_t\beta_0 + \varepsilon_t$ 得到：$\hat{\beta}_0^{ls} = \frac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t^2}$。

两边同乘 $\sqrt{z_t}$，可得 $\sqrt{z_t} y_t = \sqrt{z_t} x_t \beta_0 + \sqrt{z_t} \varepsilon_t$。

所以在 z_t 给定的条件下，$\hat{\beta}_0^{ls} = \frac{\sum_{t=1}^{T} (\sqrt{z_t} x_t)(\sqrt{z_t} y_t)}{\sum_{t=1}^{T} (\sqrt{z_t} x_t)^2}$ 是 $\sqrt{z_t} y_t = \sqrt{z_t} x_t \beta_0 + \sqrt{z_t} \varepsilon_{s_t}$ 的回归系数。

若令 $\sqrt{z_t} y_t = y_t^*$，$\sqrt{z_t} x_t = x_t^*$，$\sqrt{z_t} \varepsilon_t = \varepsilon_t^*$，则式 9.1 的最优解是 $y_t^* = x_t^* \beta_0 + \varepsilon_t^*$ 的回归系数。

二、β_1 的最优解

式 9.1 对 β_1 求一阶微分得：

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{\partial [\sum_{t=1}^{T} \ln f(y_t \mid \varphi_{t-1})]}{\partial \beta_1} = \sum_{t=1}^{T} \frac{\partial \ln f(y_t \mid \varphi_{t-1})}{\partial \beta_1}$$

$$= \sum_{t=1}^{T} \left\{ \frac{1-p}{f(y_t \mid \varphi_{t-1})} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \left[-\frac{1}{2\sigma_1^2}(y_t - x_t\beta_1) \times 2(-x_t) \right] \right\}$$

$$\Rightarrow \sum_{t=1}^{T} \frac{e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}}}{f(y_t \mid \varphi_{t-1})}(y_t - x_t\beta_1)x_t = 0$$

如令 $z_t = \dfrac{e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}}}{f(y_t \mid \varphi_{t-1})}$，则：

$$\sum_{t=1}^{T} z_t (y_t - x_t\beta_1) x_t = 0$$

$$\Rightarrow \sum_{t=1}^{T} z_t y_t x_t = \sum_{t=1}^{T} z_t x_t^2 \beta_1$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{t=1}^{T} z_t y_t x_t}{\sum_{t=1}^{T} z_t x_t^2} = \frac{\sum_{t=1}^{T} (\sqrt{z_t} x_t)(\sqrt{z_t} y_t)}{\sum_{t=1}^{T} (\sqrt{z_t} x_t)^2}$$

由 $y_t = x_t\beta_1 + \varepsilon_t$ 得到：$\hat{\beta}_1^{ls} = \dfrac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t^2}$。

两边同乘 $\sqrt{z_t}$，可得 $\sqrt{z_t} y_t = \sqrt{z_t} x_t \beta_1 + \sqrt{z_t} \varepsilon_t$。

所以在 z_t 给定的条件下，$\hat{\beta}_1^{ls} = \dfrac{\sum_{t=1}^{T} (\sqrt{z_t} x_t)(\sqrt{z_t} y_t)}{\sum_{t=1}^{T} (\sqrt{z_t} x_t)^2}$ 是 $\sqrt{z_t} y_t = \sqrt{z_t} x_t \beta_1 + \sqrt{z_t} \varepsilon_t$ 的回归系数。

若令 $\sqrt{z_t} y_t = y_t^*$，$\sqrt{z_t} x_t = x_t^*$，$\sqrt{z_t} \varepsilon_t = \varepsilon_t^*$，则式 9.1 的最优解是 $y_t^* = x_t^* \beta_1 + \varepsilon_t^*$ 的回归系数。

三、σ_0^2 的最优解

对式 9.1 求 σ_0^2 的一阶微分得：

$$\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_0^2} &= \frac{\partial \left[\sum_{t=1}^{T} \ln f(y_t \mid \varphi_{t-1}) \right]}{\partial \sigma_0^2} \\
&= \sum_{t=1}^{T} \frac{\partial \ln f(y_t \mid \varphi_{t-1})}{\partial \sigma_0^2} \\
&= \sum_{t=1}^{T} \frac{1}{f(y_t \mid \varphi_{t-1})} \frac{\partial f(y_t \mid \varphi_{t-1})}{\partial \sigma_0^2} \\
&= \sum_{t=1}^{T} \left\{ \frac{1}{f(y_t \mid \varphi_{t-1})} \frac{-2\pi (2\pi\sigma_0^2)^{-3/2}}{2} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} p + (2\pi\sigma_0^2)^{-1/2} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \right.
\end{aligned}$$

$$\left[-\frac{(y_t-x_t\beta_0)^2}{2\sigma_0^2}\right](-1)(\sigma_0^2)^{-2}p\Big\}$$

$$=\sum_{t=1}^{T}\frac{e^{-\frac{(y_t-x_t\beta_0)^2}{2\sigma_0^2}}}{f(y_t\mid\varphi_{t-1})}\left[\frac{(y_t-x_t\beta_0)^2p}{2\sqrt{2\pi}\sigma_0^5}-\frac{p}{2\sqrt{2\pi}\sigma_0^3}\right]$$

$$=0$$

如令 $z_t=\frac{e^{-\frac{(y_t-x_t\beta_0)^2}{2\sigma_0^2}}}{f(y_t\mid\varphi_{t-1})}$，则：

$\sum_{t=1}^{T}z_t[(y_t-x_t\beta_0)^2-\sigma_0^2]=0$。

$$\Rightarrow\sum_{t=1}^{T}z_t(y_t-x_t\beta_0)^2=\sum_{1}^{T}z_t\sigma_0^2$$

$$\Rightarrow\hat{\sigma}_0^2=\frac{\sum_{t=1}^{T}z_t(y_t-x_t\beta_0)^2}{\sum_{t=1}^{T}z_t}$$

四、σ_1^2 的最优解

对式 9.1 求 σ_0^2 的一阶微分得：

$$\frac{\partial\ln L}{\partial\sigma_1^2}=\frac{\partial\left[\sum_{t=1}^{T}\ln f(y_t\mid\varphi_{t-1})\right]}{\partial\sigma_1^2}$$

$$=\sum_{t=1}^{T}\frac{\partial\ln f(y_t\mid\varphi_{t-1})}{\partial\sigma_1^2}$$

$$=\sum_{t=1}^{T}\frac{1}{f(y_t\mid\varphi_{t-1})}\frac{\partial f(y_t\mid\varphi_{t-1})}{\partial\sigma_1^2}$$

$$=\sum_{t=1}^{T}\Big\{\frac{1}{f(y_t\mid\varphi_{t-1})}\frac{-2\pi(2\pi\sigma_1^2)^{-3/2}}{2}e^{-\frac{(y_t-x_t\beta_1)^2}{2\sigma_1^2}}p+(2\pi\sigma_1^2)^{-1/2}e^{-\frac{(y_t-x_t\beta_1)^2}{2\sigma_1^2}}$$

$$\left[-\frac{(y_t-x_t\beta_1)^2}{2\sigma_1^2}\right](-1)(\sigma_1^2)^{-2}p\Big\}$$

$$=\sum_{t=1}^{T}\frac{e^{-\frac{(y_t-x_t\beta_1)^2}{2\sigma_1^2}}}{f(y_t\mid\varphi_{t-1})}\left[\frac{(y_t-x_t\beta_1)^2p}{2\sqrt{2\pi}\sigma_1^5}-\frac{p}{2\sqrt{2\pi}\sigma_1^3}\right]=0$$

令 $z_t = \dfrac{e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}}}{f(y_t \mid \varphi_{t-1})}$，则：$\sum_{t=1}^{T} z_t[(y_t - x_t\beta_1)^2 - \sigma_1^2] = 0$

$$\Rightarrow \sum_{t=1}^{T} z_t (y_t - x_t\beta_1)^2 = \sum_{t=1}^{T} z_t\sigma_1^2$$

$$\Rightarrow \hat{\sigma}_1^2 = \frac{\sum_{t=1}^{T} z_t (y_t - x_t\beta_1)^2)}{\sum_{t=1}^{T} z_t}$$

五、p 的最优解

对式 9.1 求 p 的一阶微分得：

$$\frac{\partial \ln L}{\partial p} = \frac{\partial[\sum_{t=1}^{T} f(y_t \mid \varphi_{t-1})]}{\partial p} = \sum_{t=1}^{T} \frac{\partial \ln f(y_t \mid \varphi_{t-1})}{\partial p}$$

由 $p = \dfrac{e^{p_0}}{1 + e^{p_0}}$，$1 - p = 1 - \dfrac{e^{p_0}}{1 + e^{p_0}} = \dfrac{1}{1 + e^{p_0}}$，则：

$$\frac{\partial \ln L}{\partial p_0} = \sum_{t=1}^{T} \Big[\frac{1}{f(y_t \mid \varphi_{t-1})} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \frac{e^{p_0}(1 + e^{p_0}) - e^{p_0}e^{p_0}}{(1 + e^{p_0})^2}$$

$$+ \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \frac{-e^{p_0}}{(1 + e^{p_0})^2}\Big] = 0$$

$$\Rightarrow \sum_{t=1}^{T} \Big[\frac{1}{f(y_t \mid \varphi_{t-1})} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} \frac{e^{p_0}}{(1 + e^{p_0})^2} + \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}} \frac{-e^{p_0}}{(1 + e^{p_0})^2}\Big] = 0$$

$$\Rightarrow \sum_{t=1}^{T} \frac{e^{p_0}}{(1 + e^{p_0})^2} \Big[\frac{1}{f(y_t \mid \varphi_{t-1})} \Big(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_t - x_t\beta_0)^2}{2\sigma_0^2}} - \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_t - x_t\beta_1)^2}{2\sigma_1^2}}\Big)\Big] = 0$$

$$\Rightarrow \sum_{t=1}^{T} \frac{1}{f(y_t \mid \varphi_{t-1})} = 0$$

通过上式很难求出 p_0 的最优解。不仅如此，仔细观察之后不难发现我们前面所使用的权数 z_t 在计算过程中需要计算 $f(y_t \mid \varphi_{t-1})$，而该式本身就含有未知参数。所以，以上的处理办法在现实中是不可行的，需要更为复杂的处理办法来解决以上参数的估计问题。但以上这种按照一般思路来解问题的方法也给我们很多直观的认识。现有的研究中，最常用的处理办法就是我们在第二部分所讨论过的 EM 算法，采用

特殊形式——“熵”的形式——来表达似然函数。

第二节　MS - AR模型参数的EM算法

我们前面提到，如果模型存在观测值缺失或者含有不可观测变量，那么EM算法在模型极大似然函数估计中是一个可选的方法。假设该模型的未知参数 $\theta \in \Theta$，其中 Θ 为参数空间。EM 算法是一个包含“预期”（Expectation）和“最大化”（Maximization）两个步骤的不断迭代的计算过程，即：

1. 利用第（$k-1$）阶迭代中得到的参数估计值 $\theta^{(k-1)}$，我们就可以求得不可观测变量 S_t 的预期值；

2. 用此不可观测变量 S_t 的预期值代替其自身，带入似然函数，并通过最大似然法得到参数的 k 阶最优值 $\theta^{(k)}$。

每次迭代都可以改进似然函数的估计值。因此，我们可以通过给定参数初始值 $\theta^{(0)}$，不断重复上述两个步骤，一直进行到 $\theta^{(k)}$ 收敛为止。下面，我们基于 Hamilton（1989）的马尔科夫转换模型来讨论 EM 算法的具体步骤。

考虑下面两状态的 MS - AR 模型：

$$y_t = x_t\beta_{s_t} + \varepsilon_t$$

$$\varepsilon_t \sim N(0,\sigma_{s_t}^2)$$

$$\beta_{s_t} = \beta_0(1 - S_t) + \beta_1 S_t$$

$$\sigma_{s_t}^2 = \sigma_0^2(1 - S_t) + \sigma_1^2 S_t$$

$$P(S_t = 1 \mid S_{t-1} = 1) = p_{11}$$

$$P(S_t = 0 \mid S_{t-1} = 0) = p_{00}$$

假设其中的向量 x_t 由外生变量或者前定变量组成，并且 S_t 和 x_t 相互独立。我们可以将上述模型的参数分成两组 $\theta = (\theta'_1,\theta'_2)'$，第一组参数 $\theta_1 = (\beta'_0,\beta'_1,\sigma_0,\sigma_1)'$

相当于普通线性回归中的待估参数，第二组参数 $\theta_2 = (p_{00}, p_{11})'$ 为 MS－AR 模型中的状态转换概率矩阵中的参数。令 $y_1^T = (y_1, y_2, y_3, \cdots, y_T)'$，$S_1^T = (S_1, S_2, S_3, \cdots, S_T)'$，则 y_1^T 和 S_1^T 的联合密度函数以及对数似然函数可以写成：

$$
\begin{aligned}
p(y_1^T, S_1^T;\theta) &= p(y_1^T \mid S_1^T;\theta_1)p(S_1^T;\theta_2) \\
&= \prod_{t=1}^{T} p(y_t \mid S_t;\theta_1) \prod_{t=1}^{T} p(S_t \mid S_{t-1};\theta_2)
\end{aligned}
\tag{9.2}
$$

$$
\ln[p(y_1^T, S_1^T;\theta)] = \sum_{t=1}^{T} \ln[p(y_t \mid S_t;\theta_1)] + \sum_{t=1}^{T} \ln[p(S_t \mid S_{t-1};\theta_2)] \tag{9.3}
$$

如果 S_1^T 是可观测的，也就是说 S_1^T 是已知的和固定的，那么，上述似然函数最大化将与 θ_2 无关，似然函数可以只通过 θ_1 达到最大化：

$$
\frac{\partial \ln[p(y_1^T, S_1^T;\theta)]}{\partial \theta_1} = \sum_{t=1}^{T} \frac{\partial \ln[p(y_t \mid S_t;\theta_1)]}{\partial \theta_1} = 0 \tag{9.4}
$$

如果 S_1^T 不可以观测到，我们可以考虑最大化下述“熵”（Entropy）形式的似然函数：

$$
\begin{aligned}
L(\theta;y_1^T, \theta^{(k-1)}) &= \int_{S_1^T} \ln[p(y_1^T, S_1^T;\theta)]p(y_1^T, S_1^T;\theta^{(k-1)}) \\
&= \int_{S_1^T} \ln[p(y_1^T \mid S_1^T;\theta_1)p(S_1^T;\theta_2)]p(y_1^T, S_1^T;\theta^{(k-1)}) \\
&= \int_{S_1^T} [\ln p(y_1^T \mid S_1^T;\theta_1) + \ln p(S_1^T;\theta_2)]p(y_1^T, S_1^T;\theta^{(k-1)})
\end{aligned}
\tag{9.5}
$$

上式中的概率 $p(y_1^T, S_1^T;\theta^{(k-1)})$ 是在 $\theta^{(k-1)}$ 条件下得出的，并且 $\int_{S_1^T} = \sum_{S_1} \sum_{S_2} \cdots \sum_{S_T}$ 成立。

为求出关于 θ_1 的最大似然估计值，我们对参数 θ_1 求偏导，这样可以得到：

$$
\frac{\partial L(\theta;y_1^T, \theta^{(k-1)})}{\partial \theta_1} = \int_{S_1^T} \frac{\partial \ln[p(y_1^T \mid S_1^T;\theta_1)]}{\partial \theta_1} p(y_1^T, S_1^T;\theta^{(k-1)})
$$

根据贝叶斯公式 $\frac{p(A,B)}{p(A)} = p(B \mid A)$，将上式两边同时除以 $p(y_1^T;\theta^{(k-1)})$ 得到：

$$\int_{S_1^T}\frac{\partial\ln[p(y_1^T\mid S_1^T;\theta_1)]}{\partial\theta_1}\frac{p(y_1^T,S_1^T;\theta^{(k-1)})}{p(y_1^T;\theta^{(k-1)})}=0$$

$$\Rightarrow\int_{S_1^T}\frac{\partial\ln[p(y_1^T\mid S_1^T;\theta_1)]}{\partial\theta_1}p(S_1^T\mid y_1^T;\theta^{(k-1)})=0$$

根据式9.4：

$$\frac{\partial\ln[p(y_1^T,S_1^T;\theta)]}{\partial\theta_1}=\sum_{t=1}^{T}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}$$

将 y_1^T 换为 y_t 可推出：

$$\int_{S_1^T}\frac{\partial\ln[p(y_1^T\mid S_1^T;\theta_1)]}{\partial\theta_1}p(S_1^T\mid y_1^T;\theta^{(k-1)})$$

$$=\int_{S_1^T}\sum_{t=1}^{T}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}p(S_1^T\mid y_1^T;\theta^{(k-1)})$$

$$=\int_{S_t}\int_{S_{-t}}\sum_{t=1}^{T}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}p(S_1^T\mid y_1^T;\theta^{(k-1)})$$

$$=\int_{S_t}\sum_{t=1}^{T}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}\int_{S_{-t}}p(S_t,S_{-t}\mid y_1^T;\theta^{(k-1)})$$

$$=\int_{S_t}\sum_{t=1}^{T}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}p(S_t\mid y_1^T;\theta^{(k-1)})$$

所以有：

$$\sum_{t=1}^{T}\sum_{S_t=0}^{1}\frac{\partial\ln[p(y_t\mid S_t;\theta_1)]}{\partial\theta_1}p(S_t\mid y_1^T;\theta^{(k-1)})=0 \qquad (9.6)$$

由此得到的第 k 次迭代的参数估计值为 $\theta_1^{(k)}$ 。这里的 $p(S_t\mid y_1^T;\theta^{(k-1)})$ 是根据 $y_T,y_{T-1},\cdots,y_1$ 推出的平滑概率 y_t ，因此比较式9.3与式9.6可知，式9.6中 θ_1^k 是 θ_1 在 k 阶迭代所得出的最大似然估计值，并且是经过加权平均（权重为 S_t ）后，在上一期迭代值 $\theta^{(k-1)}$ 的条件下得出的平滑概率。

我们可以根据式9.6得到 $\theta_1^{(k)}=(\beta'^{(k)}_0\beta'^{(k)}_1,\sigma_0^{(k)},\sigma_1^{(k)})'$ 的具体形式。已知 $S_t=j$ ，则

$$\ln[p(y_t \mid S_t = j;\theta_1)] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(\ln\sigma_j^2) - \frac{1}{2}\frac{(y_t - x_t\beta_j)^2}{\sigma_j^2}$$

将上式代入式9.5，并对参数 β_j 求导得：

$$\sum_{t=1}^{T}\sum_{S_t=0}^{1}\frac{\partial\ln[p(y_t \mid S_t)]}{\partial\beta_j}p(S_t \mid y_1^T;\theta^{(k-1)})$$

$$= \sum_{t=1}^{T}\frac{x_t}{\sigma_j^2}(y_t - x_t\beta_j)p(S_t = j \mid y_1^T;\theta^{(k-1)}) = 0$$

同理对 σ_j^2 求导得：

$$\sum_{t=1}^{T}\sum_{S_t=0}^{1}\frac{\partial\ln[p(y_t \mid S_t)]}{\partial\sigma_j^2}p(S_t \mid y_1^T;\theta^{(k-1)})$$

$$\sum_{t=1}^{T}\left\{-\frac{1}{2\sigma_j^2} + \frac{1}{2}\frac{(y_t - x_t\beta_j)^2}{\sigma_j^4}\right\}p(S_t = j \mid y_1^T;\theta^{(k-1)}) = 0$$

由此可以得出 $\beta_j^{(k)}$ 和 $\sigma_j^{(k)}$ ：

$$\beta_j^{(k)} = \frac{\sum_{t=1}^{T}x_t y_t p(S_t = j \mid y_1^T;\theta^{(k-1)})}{\sum_{t=1}^{T}x_t x'_t p(S_t = j \mid y_1^T;\theta^{(k-1)})}$$

$$= \frac{\sum_{t=1}^{T}[x_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}][y_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}]}{\sum_{t=1}^{T}(x_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})})^2} \quad j = 0,1 \tag{9.7}$$

$$\sigma_j^{(k)} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - x_t\beta_j^{(k)})^2 p(S_t = j \mid y_1^T;\theta^{(k-1)})}{\sum_{t=1}^{T}p(S_t = j \mid y_1^T;\theta^{(k-1)})}} \quad j = 0,1 \tag{9.8}$$

通过式9.7和式9.8可以看出 $\beta_j^{(k)}$ 是 $\tilde{y}_t = \beta_j^{(k)}\tilde{x}_t + \varepsilon_t$ 经过回归所得的参数，其中：

$$\tilde{y}_t = y_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}$$

$$\tilde{x}_t = x_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}$$

同样，采用类似方法，如式9.5和式9.6，可知：

$$p_{ij}^{(k)} = \frac{\sum_{t=1}^{T} p(S_t = j, S_{t-1} = i \mid y_1^T; \theta^{(k-1)})}{\sum_{t=1}^{T} p(S_{t-1} = i \mid y_1^T; \theta^{(k-1)})} \qquad j = 0,1\ ;\ \sum_{j=0}^{1} p_{ij}^{(k)} = 1$$

EM 方法的一个优点在于在求极大值的步骤中很容易得出参数的估计值，不需要很复杂的求最优解的过程，并且这种方法对于参数初始值选择来讲具有稳健性。

第三节　MS - AR (1) 模型的详细计算过程：Excel 应用

对于简单的回归模型：$y_i = \alpha + \beta x_i + \varepsilon_i$，采用最小二乘法可以求出参数 $\hat{\alpha}_{ls}$ 和 $\hat{\beta}_{ls}$ 的值，即：

$$\begin{cases} \hat{\beta}_{ls} = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - x)^2} \\ \hat{\alpha}_{ls} = \bar{y} - \hat{\beta}_{ls}\bar{x} \end{cases}$$

通过 Excel 进行回归，求出 $\hat{\alpha}_{ls}$ 和 $\hat{\beta}_{ls}$ 的步骤如表 9 - 1。

表 9 - 1

y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x - \bar{x})$	$(x_i - \bar{x})^2$
y_1	x_1	$(y_1 - \bar{y})$	$(x_1 - \bar{x})$	$(y_1 - \bar{y})(x_1 - \bar{x})$	$(x_i - \bar{x})^2$
y_2	x_2	$(y_2 - \bar{y})$	$(x_2 - \bar{x})$	$(y_2 - \bar{y})(x_2 - \bar{x})$	$(x_i - \bar{x})^2$
…	…	…	…	…	…
y_n	x_n	$(y_n - \bar{y})$	$(x_n - \bar{x})$	$(y_n - \bar{y})(x_n - \bar{x})$	$(x_n - \bar{x})^2$
$\sum y_i$	$\sum x_i$	—	—	$A = \sum (y_i - \bar{y})(x_i - \bar{x})$	$B = \sum (x_i - \bar{x})^2$
$\bar{y}$	$\bar{x}$	—	—	—	—

根据表中的内容，可以求出 $\hat{\beta}_{ls} = \dfrac{A}{B}$，$\hat{\alpha}_{ls} = \bar{y} - \hat{\beta}_{ls}\bar{x}$。

一、MS－AR（1）模型的具体算法回顾

1. 初始值

一阶自回归模型：$y_t = \alpha_{S_t} + \beta y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim i.i.d.N(0,\sigma^2)$

可以转化为：$y_t - \bar{y}_{S_t} = \beta(y_{t-1} - \bar{y}_{S_{t-1}}) + v_t \quad v_t \sim i.i.d.N(0,\sigma^2_{S_{t-1}})$

参数包括：$\beta,\bar{y}_0,\bar{y}_1,\sigma_0^2,\sigma_1^2$。其中 $\bar{y}_0,\bar{y}_1,\sigma_0^2,\sigma_1^2$ 决定正态分布的参数；马尔科夫转换矩阵的参数为 p_{00},p_{11}；$\delta = [P(S_t = 0),P(S_t = 1)]'$ 是稳定状态分布。

2. 过滤过程

预测过程：

$$f(y_t \mid y_1^{t-1}) = \sum_{S_t=1}^{M}\sum_{S_{t-1}=1}^{M} f(y_t,S_t,S_{t-1} \mid y_1^{t-1})$$

$$= \sum_{S_t=1}^{M}\sum_{S_{t-1}=1}^{M} f(y_t \mid S_t,S_{t-1},y_1^{t-1})P(S_t,S_{t-1} \mid y_1^{t-1})$$

更新过程：

$$P(S_t = j,S_{t-1} = i \mid y_1^t)$$

$$= \frac{f(y_t \mid S_t = j,S_{t-1} = i,y_1^{t-1})P(S_t = j,S_{t-1} = i \mid y_1^{t-1})}{\sum_{S_t=1}^{M}\sum_{S_{t-1}=1}^{M} f(y_t \mid S_t = j,S_{t-1} = i,y_1^{t-1})P(S_t = j,S_{t-1} = i \mid y_1^{t-1})}$$

得出：$P(S_t = j \mid y_1^t) = \sum_{S_{t-1}=1}^{M} P(S_t = j,S_{t-1} = i \mid y_1^t)$

3. 平滑过程：

$$P(S_t = j \mid y_1^T) = \sum_{k=1}^{M} P(S_t = j,S_{t+1} = k \mid y_1^T)$$

4. 估计过程：EM 算法

$$\beta_j^{(k)} = \frac{\sum_{t=1}^{T}[x_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}][y_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})}]}{\sum_{t=1}^{T}(x_t\sqrt{p(S_t = j \mid y_1^T;\theta^{(k-1)})})^2}$$

$$\sigma_j^{(k)} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - x_t\beta_j^{(k)})^2 p(S_t = j \mid y_1^T;\theta^{(k-1)})}{\sum_{t=1}^{T} p(S_t = j \mid y_1^T;\theta^{(k-1)})}}$$

$$p_{ij}^{(k)} = \frac{\sum_{t=1}^{T} p(S_t = j, S_{t-1} = i \mid y_1^T; \theta^{(k-1)})}{\sum_{t=1}^{T} p(S_{t-1} = i \mid y_1^T; \theta^{(k-1)})}$$

5. 判断收敛

$$\begin{cases} \text{如果收敛，即得出结果；} \\ \text{否则，执行第 6 步。} \end{cases}$$

6. 更新初始值，返回 2。

二、MS－AR（1）模型算法在 Excel 中的应用过程

1. 初始值的设定：$\beta, \bar{y}_0, \bar{y}_1, \sigma_0^2, \sigma_1^2, p_{00}, p_{11}$，并据此计算出：

$$\delta = \begin{bmatrix} \dfrac{1 - p_{11}}{2 - p_{00} - p_{11}} \\ \\ \dfrac{1 - p_{00}}{2 - p_{00} - p_{11}} \end{bmatrix} = \begin{bmatrix} \delta_0 \\ \\ \delta_1 \end{bmatrix}$$

2. 过滤过程：F1—F21

（1）求 $P(S_t = 0 \mid y_1^{t-1})$ 和 $P(S_t = 1 \mid y_1^{t-1})$ 的：F1—F6

（2）求 $f(y_t \mid y_1^{t-1})$：F7—F8

（3）更新数据：求 $P(S_t \mid y_1^T)$：F9—F21

3. 平滑过程：A1—A2

初始值：$P(S_T = 0 \mid y_1^T)$ 和 $P(S_T = 1 \mid y_1^T)$

求 $P(S_t = 0 \mid y_1^T)$ 和 $P(S_t = 1 \mid y_1^T)$：A1—A6

4. 估计过程：E1—E4

5. 判断收敛：

若满足收敛条件：$\dfrac{L_k - L_{k-1}}{L_{k-1}} \leq 0.01$，则得出结果；否则进入第 6 步。

6. 计算出参数值 $\beta, \bar{y}_0, \bar{y}_1, \sigma_0^2, \sigma_1^2, p_{00}, p_{11}$，重复 2 到 6。

三、Excel 表格计算过程

1. 过滤过程一：求解 $P(S_t=1\mid y_1^{t-1})$ $P(S_t=0\mid y_1^{t-1})$。

表 9－1

	F1	F2	F3	F4
t	$P(S_t=0,S_{t-1}=0\mid y_1^{t-1})$ $=P(S_t=0\mid S_{t-1}=0)P(S_{t-1}=0\mid y_1^{t-1})$	$P(S_t=0,S_{t-1}=1\mid y_1^{t-1})$ $=P(S_t=0\mid S_{t-1}=1)P(S_{t-1}=1\mid y_1^{t-1})$	$P(S_t=1,S_{t-1}=0\mid y_1^{t-1})$ $=P(S_t=1\mid S_{t-1}=0)P(S_{t-1}=0\mid y_1^{t-1})$	$P(S_t=1,S_{t-1}=1\mid y_1^{t-1})$ $=P(S_t=1\mid S_{t-1}=1)P(S_{t-1}=1\mid y_1^{t-1})$
1	$P(S_1=0,S_0=0\mid y_0)$ $=P(S_1=0\mid s_0=0)P(S_0=0\mid y_0)$	$P(S_1=0,S_0=1\mid y_0)$ $=P(S_1=0\mid S_0=1)P(S_0=1\mid y_0)$	$P(S_1=1,S_0=0\mid y_0)$ $=P(S_1=1\mid S_0=0)P(S_0=0\mid y_0)$	$P(S_1=1,S_0=1\mid y_0)$ $=P(S_1=1\mid S_0=1)P(S_0=1\mid y_0)$
…	…	…	…	…
T	$P(S_t=1,S_{t-1}=0\mid y_1^{T-1})$ $=P(S_t=1\mid S_{t-1}=0)P(S_{t-1}=0\mid y_1^{T-1})$	$P(S_t=0,S_{t-1}=1\mid y_1^{T-1})$ $=P(S_t=0\mid S_{t-1}=1)P(S_{t-1}=1\mid y_1^{T-1})$	$P(S_t=1,S_{t-1}=0\mid y_1^{T-1})$ $=P(S_t=1\mid S_{t-1}=0)P(S_{t-1}=0\mid y_1^{T-1})$	$P(S_t=1,S_{t-1}=1\mid y_1^{T-1})$ $=P(S_t=1\mid S_{t-1}=1)P(S_{t-1}=1\mid y_1^{T-1})$

	F5 = F1 + F2	F6 = F3 + F4
t	$P(S_t=0\mid y_1^{t-1})=P(S_t=0,S_{t-1}=0\mid y_1^{t-1})+P(S_t=0,S_{t-1}=1\mid y_1^{t-1})$	$P(S_t=1\mid y_1^{t-1})=P(S_t=1,S_{t-1}=0\mid y_1^{t-1})+P(S_t=1,S_{t-1}=1\mid y_1^{t-1})$
1	$P(S_1=0\mid y_0)=P(S_1=0,S_0=0\mid y_0)+P(S_1=0,S_0=1\mid y_0)$	$P(S_1=1\mid y_0)=P(S_1=1,S_0=0\mid y_0)+P(S_1=1,S_0=1\mid y_0)$
…	…	…
T	$P(S_T=0\mid y_1^{T-1})=P(S_T=0,S_{T-1}=0\mid y_1^{T-1})+P(S_T=0,S_{T-1}=1\mid y_1^{T-1})$	$P(S_T=1\mid y_1^{T-1})=P(S_T=1,S_{T-1}=0\mid y_1^{T-1})+P(S_T=1,S_{T-1}=1\mid y_1^{T-1})$

2. 过滤过程二：求解 $f(y_t \mid S_t=0)\,f(y_t \mid S_t=1)$。

表 9-2

	F7.1	F7.2	F7 = F7.1 + F7.2
t	$f(y_t \mid s_t=0, s_{t-1}=0)$	$f(y_t \mid s_t=0, s_{t-1}=1)$	$f(y_t \mid s_t=0)$
1	$f(y_1 \mid S_1=0, S_0=0) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_1-\bar{y}_0)-\beta(y_0-\bar{y}_0)]^2}{2\sigma_0^2}\right\}$	$f(y_1 \mid S_1=0, S_0=1) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_1-\bar{y}_0)-\beta(y_0-\bar{y}_1)]^2}{2\sigma_0^2}\right\}$	$f(y_1 \mid S_1=0) = \sum_{S_t=0}^{1}\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_1-\bar{y}_0)-\beta(y_0-\bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}$
…	…	…	…
T	$f(y_T \mid S_T=0, S_{T-1}=0) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_T-\bar{y}_0)-\beta(y_{T-1}-\bar{y}_0)]^2}{2\sigma_0^2}\right\}$	$f(y_T \mid S_T=0, S_{T-1}=0) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_T-\bar{y}_0)-\beta(y_{T-1}-\bar{y}_1)]^2}{2\sigma_0^2}\right\}$	$f(y_T \mid S_T=0) = \sum_{S_t=0}^{1}\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_T-\bar{y}_0)-\beta(y_{T-1}-\bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}$
	F8.1	F8.2	F8 = F8.1 + F8.2
t	$f(y_t \mid s_t=1, s_{t-1}=0)$	$f(y_t \mid s_t=1, s_{t-1}=1)$	$f(y_t \mid s_t=1)$
1	$f(y_1 \mid S_1=1, S_0=0) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_1-\bar{y}_1)-\beta(y_0-\bar{y}_0)]^2}{2\sigma_1^2}\right\}$	$f(y_1 \mid S_1=1, S_0=1) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_1-\bar{y}_1)-\beta(y_0-\bar{y}_1)]^2}{2\sigma_1^2}\right\}$	$f(y_1 \mid S_1=1) = \sum_{S_t=0}^{1}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_1-\bar{y}_1)-\beta(y_0-\bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}$
…	…	…	…
T	$f(y_T \mid S_T=1, S_{T-1}=0) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_T-\bar{y}_0)-\beta(y_{T-1}-\bar{y}_0)]^2}{2\sigma_1^2}\right\}$	$f(y_T \mid S_T=1, S_{T-1}=1) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_T-\bar{y}_1)-\beta(y_{T-1}-\bar{y}_1)]^2}{2\sigma_1^2}\right\}$	$f(y_T \mid S_T=1) = \sum_{S_t=0}^{1}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_T-\bar{y}_1)-\beta(y_{T-1}-\bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}$

续表

	F9 = F7 * F5	F10 = F8 * F6	F11 = F9 + F10
t	$f(y_t \mid S_t = 0)P(S_t = 0 \mid y_1^{t-1})$	$f(y_t \mid S_t = 1)P(S_t = 1 \mid y_1^{t-1})$	$f(y_t \mid y_1^{t-1})$
1	$f(y_1 \mid S_1 = 0)P(S_1 = 0 \mid y_0)$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp$ $\left\{-\frac{[(y_1 - \bar{y}_0) - \beta(y_0 - \bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}\cdot$ $[P(S_1 = 0, S_1 = 0 \mid y_0) + P(S_1 = 0, S_1 = 1 \mid y_0)]$	$f(y_1 \mid S_1 = 1)P(S_1 = 1 \mid y_0)$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp$ $\left\{-\frac{[(y_1 - \bar{y}_1) - \beta(y_0 - \bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}\cdot$ $[P(S_1 = 1, S_1 = 0 \mid y_0) + P(S_1 = 1, S_1 = 1 \mid y_0)]$	$f(y_t \mid y_1^{t-1})$ $= f(y_1 \mid S_1 = 0)P(S_1 = 0 \mid y_0) + f(y_1 \mid S_1 = 1)P(S_1 = 1 \mid y_0)$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_1 - \bar{y}_0) - \beta(y_0 - \bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}\cdot$ $[P(S_1 = 0, S_0 = 0 \mid y_0) + P(S_1 = 0, S_0 = 1 \mid y_0)]$ $+ \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_1 - \bar{y}_1) - \beta(y_0 - \bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}\cdot$ $[P(S_1 = 1, S_0 = 0 \mid y_0) + P(S_1 = 1, S_0 = 1 \mid y_0)]$
…	…	…	…
T	$f(y_T \mid S_T = 0)P(S_T = 0 \mid y_{T-1})$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp$ $\left\{-\frac{[(y_T - \bar{y}_0) - \beta(y_{T-1} - \bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}$ $\cdot [P(S_T = 0, S_{T-1} = 0 \mid y_{T-1}) + P(S_T = 0, S_{T-1} = 1 \mid y_{T-1})]$	$f(y_T \mid S_T = 1)P(S_T = 1 \mid y_{T-1})$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp$ $\left\{-\frac{[(y_T - \bar{y}_0) - \beta(y_{T-1} - \bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}$ $\cdot [P(S_T = 1, S_{T-1} = 0 \mid y_{T-1}) + P(S_T = 1, S_{T-1} = 1 \mid y_{T-1})]$	$f(y_T \mid y_1^{T-1})$ $= f(y_T \mid S_T = 0)P(S_T = 0 \mid y_1^{T-1}) + f(y_T \mid S_T = 1)P(S_T = 1 \mid y_1^{T-1})$ $= \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{[(y_T - \bar{y}_0) - \beta(y_{T-1} - \bar{y}_{S_0})]^2}{2\sigma_0^2}\right\}$ $\cdot [P(S_T = 0, S_{T-1} = 0 \mid y_0) + P(S_T = 0, S_{T-1} = 1 \mid y_0)]$ $+ \sum_{S_0=0}^{1} \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{[(y_1 - \bar{y}_1) - \beta(y_0 - \bar{y}_{S_0})]^2}{2\sigma_1^2}\right\}\cdot$ $[P(S_T = 1, S_{T-1} = 0 \mid y_0) + P(S_T = 1, S_{T-1} = 1 \mid y_0)]$

3. 过滤过程三：更新 $P(S_t \mid y_1^t)$。

表 9－3

F_{12}	F_{13}	F_{14}	F_{15}
$f(y_t, S_t = 0, S_{t-1} = 0 \mid y_1^{t-1})$	$f(y_t, S_t = 0, S_{t-1} = 1 \mid y_1^{t-1}))$	$f(y_t, S_t = 1, S_{t-1} = 0 \mid y_1^{t-1}))$	$f(y_t, S_t = 1, S_{t-1} = 1 \mid y_1^{t-1}))$
$= f(y_t \mid S_t = 0, S_{t-1} = 0) \cdot P(S_t = 0, S_{t-1} = 0 \mid y_1^{t-1})$	$= f(y_t \mid S_t = 0, S_{t-1} = 1) \cdot P(S_t = 0, S_{t-1} = 1 \mid y_1^{t-1})$	$= f(y_t \mid S_t = 1, S_{t-1} = 0) \cdot P(S_t = 1, S_{t-1} = 0 \mid y_1^{t-1})$	$= f(y_t \mid S_t = 1, S_{t-1} = 1) \cdot P(S_t = 1, S_{t-1} = 1 \mid y_1^{t-1})$
$= F_{7.1} * F_1$	$= F_{7.2} * F_2$	$= F_{8.1} * F_3$	$= F_{8.2} * F_4$

F_{16}	F_{17}	F_{18}	F_{19}
$f(y_t, S_t = 0, S_{t-1} = 0 \mid y_1^t)$	$f(y_t, S_t = 0, S_{t-1} = 1 \mid y_1^t)$	$f(y_t, S_t = 1, S_{t-1} = 0 \mid y_1^t)$	$f(y_t, S_t = 1, S_{t-1} = 1 \mid y_1^t)$
$= \frac{f(y_t, S_t = 0, S_{t-1} = 0 \mid y_1^{t-1})}{f(y_t \mid y_1^{t-1})}$	$= \frac{f(y_t, S_t = 0, S_{t-1} = 1 \mid y_1^{t-1})}{f(y_t \mid y_1^{t-1})}$	$= \frac{f(y_t, S_t = 1, S_{t-1} = 0 \mid y_1^{t-1})}{f(y_t \mid y_1^{t-1})}$	$= \frac{f(y_t, S_t = 1, S_{t-1} = 1 \mid y_1^{t-1})}{f(y_t \mid y_1^{t-1})}$
$= \frac{F_{12}}{F_{11}}$	$= \frac{F_{13}}{F_{11}}$	$= \frac{F_{14}}{F_{11}}$	$= \frac{F_{15}}{F_{11}}$

F_{20}	F_{21}
$P(S_t = 0 \mid y_1^t)$	$P(S_t = 1 \mid y_1^t)$
$= P(S_t = 0, S_{t-1} = 0 \mid y_1^t) + P(S_t = 0, S_{t-1} = 1 \mid y_1^t)$	$= P(S_t = 1, S_{t-1} = 0 \mid y_1^t) + P(S_t = 1, S_{t-1} = 1 \mid y_1^t)$
$= F_{16} + F_{17}$	$= F_{18} + F_{19}$

4. 平滑过程：$P(S_t = 0 \mid y_1^T)$ $P(S_t = 1 \mid y_1^T)$。

表 9－4

A_1	A_2
$P(S_t = 0, S_{t-1} = 0 \mid y_1^T)$	$P(S_t = 0, S_{t-1} = 1 \mid y_1^T)$
$= \frac{P(S_{t+1} = 0 \mid y_1^T)P(S_t = 0 \mid y_1^t)P(S_{t+1} = 0 \mid S_t = 0)}{P(S_{t+1} = 0 \mid y_1^t)}$	$= \frac{P(S_{t+1} = 0 \mid y_1^T)P(S_t = 1 \mid y_1^t)P(S_{t+1} = 0 \mid S_t = 1)}{P(S_{t+1} = 0 \mid y_1^t)}$
$= \frac{P(S_{t+1} = 0 \mid y_1^T)F_{20}p_{00}}{F_5}$	$= \frac{P(S_{t+1} = 0 \mid y_1^T)F_{21}(1 - p_{11})}{F_5}$

A_3	A_4
$P(S_t = 1, S_{t-1} = 0 \mid y_1^T)$	$P(S_t = 1, S_{t-1} = 1 \mid y_1^T)$
$= \frac{P(S_{t+1} = 1 \mid y_1^T)P(S_t = 0 \mid y_1^t)P(S_{t+1} = 1 \mid S_t = 0)}{P(S_{t+1} = 1 \mid y_1^t)}$	$= \frac{P(S_{t+1} = 1 \mid y_1^T)P(S_t = 1 \mid y_1^t)P(S_{t+1} = 1 \mid S_t = 1)}{P(S_{t+1} = 1 \mid y_1^t)}$
$= \frac{P(S_{t+1} = 1 \mid y_1^T)F_{20}p_{00}}{F_6}$	$= \frac{P(S_{t+1} = 1 \mid y_1^T)F_{21}p_{11}}{F_6}$

A_5	A_6
$P(S_t = 0 \mid y_1^T)$	$P(S_t = 1 \mid y_1^T)$
$= P(S_t = 0, S_{t-1} = 0 \mid y_1^T) + P(S_t = 0, S_{t-1} = 1 \mid y_1^T)$	$= P(S_t = 1, S_{t-1} = 0 \mid y_1^T) + P(S_t = 1, S_{t-1} = 1 \mid y_1^T)$
$= A_1 + A_2$	$= A_3 + A_4$

5. 估计过程：EM 算法：

表 9－5

		E1	E2	E3	E4
		$\tilde{y}_{0,t}$	$\tilde{y}_{1,t}$	$\tilde{x}_{0,t}$	$\tilde{x}_{1,t}$
y_t	x_t	$y_t\sqrt{P[S_t=0\mid y_1^t]}$	$y_t\sqrt{P[S_t=1\mid y_1^t]}$	$x_t\sqrt{P[S_t=0\mid y_1^t]}$	$x_t\sqrt{P[S_t=1\mid y_1^t]}$
y_1	x_1	$y_1\sqrt{P[S_t=0\mid y_1^t]}$	$y_1\sqrt{P[S_t=1\mid y_1^t]}$	$x_1\sqrt{P[S_t=0\mid y_1^t]}$	$x_1\sqrt{P[S_t=1\mid y_1^t]}$
…	…	…	…	…	…
y_T	y_T	$y_T\sqrt{P[S_t=0\mid y_1^t]}$	$y_T\sqrt{P[S_t=1\mid y_1^t]}$	$x_T\sqrt{P[S_t=0\mid y_1^t]}$	$x_T\sqrt{P[S_t=1\mid y_1^t]}$
		$y_i\sqrt{A_5}$	$y_i\sqrt{A_6}$	$x_i\sqrt{A_5}$	$x_i\sqrt{A_6}$

由 EM 算法可知，通过 $\tilde{y}_{0,t}$ 对 $\tilde{x}_{0,t}$ 回归可以得到β_0，$\tilde{y}_{1,t}$ 对 $\tilde{x}_{1,t}$ 回归可以得到β_1。这样，我们根据前面的运算结果可以求出：

$$\hat{\beta}_0=\frac{\tilde{y}_{0,t}\tilde{x}_{0,t}}{\tilde{x}_{0,t}^2}=\frac{E_1\times E_3\times A_5}{E_3^2},\ \hat{\beta}_1=\frac{\tilde{y}_{1,t}\tilde{x}_{1,t}}{\tilde{x}_{1,t}^2}=\frac{E_2\times E_4\times A_6}{E_4^2}$$

由 EM 算法又知：

$$\sigma_j^{(k)}=\sqrt{\frac{\sum_{j=0}^{1}[y_t-\bar{y}_{S_t}-\beta^{(k)}(y_{t-1}-\bar{y}_{S_{t-1}})]^2p(S_t=j\mid\tilde{y}_{j,1}^T;\theta^{(k-1)})}{\sum_{j=0}^{1}p(S_t=j\mid\tilde{y}_{j,1}^T;\theta^{(k-1)})}}$$

，因此可以得出 $\hat{\sigma}_0$ 和 $\hat{\sigma}_1$：

$$\sigma_0^{(k)}=\sqrt{\frac{[y_t-\bar{y}_{S_t}-\beta^{(k)}(y_{t-1}-\bar{y}_{S_{t-1}})]^2p(S_t=0\mid\tilde{y}_{0,1}^T)}{\sum_{j=0}^{1}p(S_t=j\mid\tilde{y}_{j,1}^T)}}$$

$$\sigma_1^{(k)}=\sqrt{\frac{[y_t-\bar{y}_{S_t}-\beta^{(k)}(y_{t-1}-\bar{y}_{S_{t-1}})]^2p(S_t=1\mid\tilde{y}_{0,1}^T)}{\sum_{j=0}^{1}p(S_t=j\mid\tilde{y}_{j,1}^T)}}$$

同样，我们由 EM 算法可知：

$$p_{ij}^{(k)}=\frac{\sum_{t=1}^{T}p(S_t=j,S_{t-1}=j\mid\tilde{y}_{j,1}^T;\theta^{(k-1)})}{\sum_{t=1}^{T}p(S_{t-1}=j\mid\tilde{y}_{j,1}^T;\theta^{(k-1)})}$$

，因此，可以得出 p_{00} 和 p_{11}：

$$p_{00}=P(S_0=0\mid y_0)=\frac{1-p}{2-p-q},\ p_{11}=P(S_0=1\mid y_0)=\frac{1-p}{2-p-q}$$

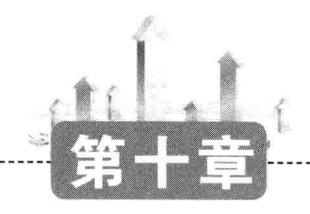
第四部分

HMM 和 MS－AR 模型应用

以上三部分内容讲解了 HMM 和 MS - AR 模型的设置、估计、预测等理论内容，本部分将介绍这些理论模型在宏观经济研究、金融投资研究等实际操作领域中的应用。

金融投资不能脱离具体的宏观经济形势，需要以宏观基本面为基础选择投资时机和投资策略。所以，第十章首先以中国和美国的 GDP 数据为例，介绍如何将 MS - AR 模型应用于宏观经济研究领域。具体来说，即如何判断过去和将来经济运行的方向及概率分布。Hamilton（1989）不仅是 MS - AR 模型在宏观经济领域研究的典范，而且也对后面金融领域市场阶段的划分起到很强的借鉴作用。所以，该章将 MS - AR 模型应用于中国和美国宏观经济研究，并介绍 Kim 和 Nelson（1999）对该模型做出了一些贡献。

然后，第十一章将介绍如何把 HMM 和 MS - AR 模型应用于我国股票市场研究。金融市场的特征之一就是均值和方差会随时间和市场形态而发生变化。根据这些特点，这部分内容可以帮助我们解决两个问题：股票市场中的牛市和熊市如何划分；如何判断市场风险的阶段变化。解决第一个问题需要利用 HMM，考虑无序列相关或有序列直接相关的均值转换模型；第二个问题需要考虑带方差转换的 SWARCH 模型。我们希望通过这几个模型的建立和求解，能够了解我国股票市场运行的基本特征，为投资者提供可以获取“超额信息”的可靠方法，为后续研究奠定坚实的基础。



第十章 MS－AR模型在宏观经济分析中的应用

第一节 简单MS-AR (1) 经济波动模型

在用MS－AR模型研究经济周期或经济波动时，转折点（Turning Point）的识别对于宏观经济政策具有很重大的意义。转折点代表数据中固有的结构性变化，从而将前后两个阶段的数据显著地区分开。这种模型的一个重要特点在于：对经济波动中的非线性和不对称性特征能够很好地描述。例如Hamilton（1989）假设经济周期分为衰退和扩张两个时期，从而利用两状态马尔科夫转换模型描述实际GNP增长率。

下面，我们将运用简单的一阶自相关马尔科夫转换模型，即MR－AR（1）模型，对我国的GDP数据进行分析，以下采用中国1999—2012年GDP季度数据。根据Harvey（1989）的研究结论，时间序列结构化模型要考虑变量的季节性成分、趋势性成分、周期性成分等。所以，首先将水平数据去掉季节性因素，再转化为增长率数据，以去掉趋势性成分，那么经过处理数据可以视为只含周期性成分和随机成分的序列。将得到的时间序列记为Δy_t，考虑下面简单的变均值的MS－AR（1）模型：

$$(\Delta y_t - \mu_{S_t}) = \varphi_1(\Delta y_{t-1} - \mu_{S_{t-1}}) + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N(0,\sigma^2) \qquad (10.1)$$

$$\mu_{S_t} = \mu_0(1 - S_t) + \mu_1 S_t \qquad (10.2)$$

$$P(S_t = 1 \mid S_{t-1} = 1) = p \qquad (10.3)$$

$$P(S_t = 0 \mid S_{t-1} = 0) = q \tag{10.4}$$

用 Matlab 程序进行处理，平滑概率的计算结果如下：

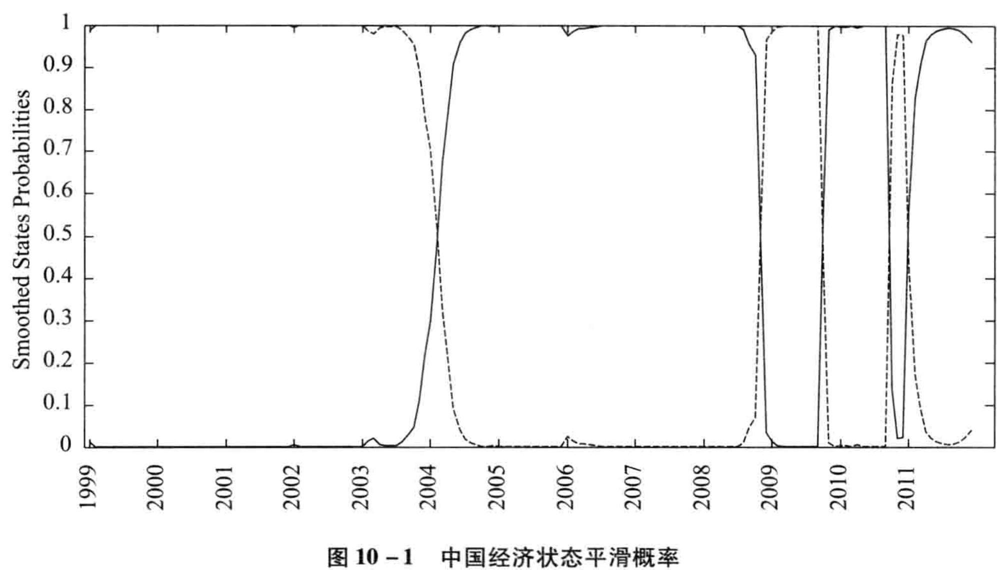


图 10－1　中国经济状态平滑概率

图 10－1 中状态 1 的均值较大，所以状态 1 对应着经济周期中的膨胀状态，在图中用实线表示；状态 0 的均值较小，对应着衰退状态，在图中用虚线表示。从中国季度 GDP 数据的分析结果中不难看出，在 1999—2003 年，中国经济处于较低的增长区间，从 2004 年至 2008 年下半年中国经济保持了较高的增长率，而在 2009 年之后经济运行缺乏稳定性，经济波动性加大。利用这一部分样本可以发现，状态 0 和状态 1 呈现交替出现的态势，两个状态的转移矩阵为：

$$\begin{pmatrix} 0.97 & 0.03 \\ 0.04 & 0.96 \end{pmatrix}$$

所以，从理论上来讲，状态 1 持续期大致为 1/（1－0.97）＝33 个季度，状态 2 持续期稍短，大致为 1/（1－0.96）＝25 个季度。

同理，利用美国同期 GDP 季度数据进行分析得到平滑概率计算结果如图 10－2 所示：

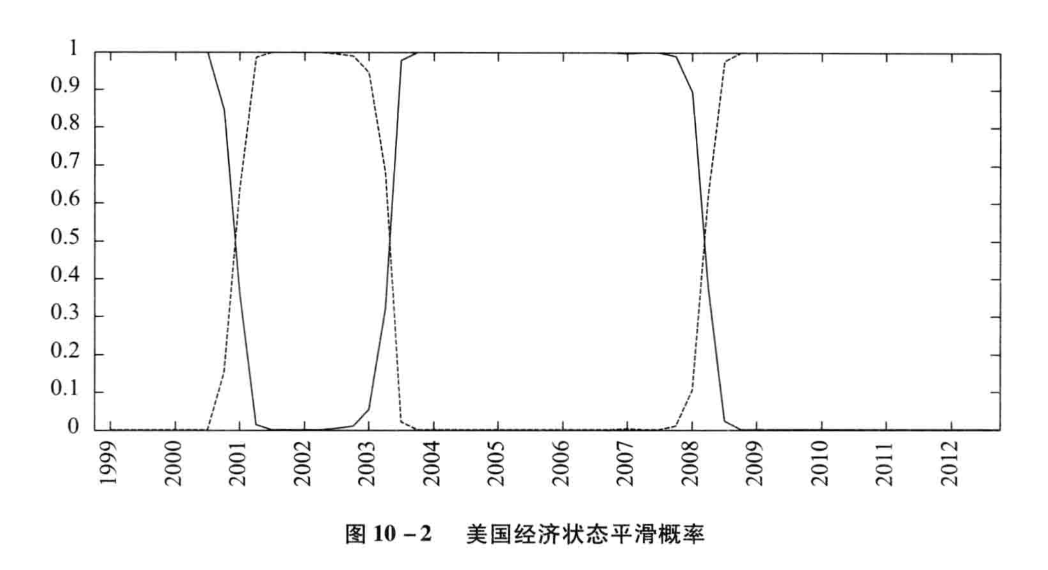


图 10－2　美国经济状态平滑概率

参照前面对中国经济分析的思路，图中状态 1 对应着经济周期中的膨胀状态，用实线表示；状态 0 为经济周期的衰退阶段，用虚线表示。美国在 2000 年下半年到 2003 年上半年处于经济周期的衰退阶段，之后经济好转并保持了较高的增长率。但是在 2008 年上半年重新落入衰退阶段，至今并未有明显迹象表明美国经济已经处于膨胀阶段。这与美国国家经济研究局（NBER[①]）发布的官方经济周期是基本一致的。根据 NBER 公布的数据显示，美国经济在 2001 年上半年开始出现衰退趋势，并在 2001 年 11 月达到谷底，之后于 2007 年 10 月开始再次滑向谷底。

美国经济状态转移矩阵与中国经济状态转移矩阵相同，持续期也都相同。在过去的 15 年间，作为最大的发达国家和最大的发展中国家，美国和中国经济体现出周期长度的基本一致性。但是，这并不意味着中美两国在经济周期问题上具有同步性，从图中也可以发现，两者所经历的经济周期各个阶段在时间点上有所差别[②]。

正如本节开始所言，这是一个简单的 MS－AR（1）模型，所以得到的结论也只是为了说明模型的使用方法和实际应用中的基本效果，还有很大的改进余地。MS－

① http：//www. nber. org。
② 这是由两国发展中的客观实际决定的，包括发展模式和发展阶段等。目前来看，两国均未摆脱经济发展中的衰退周期，但也许是因为增长趋势的永久变化，这里不展开讨论。

AR（1）模型没有 Hamilton（1989）的 MS – AR（4）模型复杂，但有趣的是，二者计算结果很接近。下面，我们将详细讨论 Hamilton（1989）的 MS – AR（4）模型。该模型不仅在宏观经济分析中得到广泛采用，在金融市场分析中也发挥着重要的作用。

第二节　Hamilton(1989)和 Kim, Nelson (1999) 的 MS - AR (4) 经济波动模型

本节将重点讨论 Hamilton（1989）经济波动的 MS – AR（4）模型。Hamilton（1989）将马尔科夫转换模型应用到经济波动的分析中，着重强调转折点的内生性。这类模型的重要特征就是它们能够很好地描述经济波动中非线性动态或者不对称性因素。例如，在 Hamilton（1989）的两状态马尔科夫转换过程中，我们可以通过 GNP 增长率所处的不同状态，来区别出衰退与膨胀两种经济状态的动态变化。Hamilton（1989）的 MS – AR（4）模型如下：

$$
\begin{aligned}
(\Delta y_t - \mu_{s_t}) &= \varphi_1(\Delta y_{t-1} - \mu_{s_{t-1}}) + \varphi_2(\Delta y_{t-2} - \mu_{s_{t-2}}) + \cdots \\
&\quad + \varphi_4(\Delta y_{t-4} - \mu_{s_{t-4}}) + \varepsilon_t \text{ , } \varepsilon_t \sim i.i.d.\ N(0,\sigma^2) \qquad (10.5) \\
\mu_{s_t} &= \mu_0(1 - S_t) + \mu_1 S_t \\
P(S_t &= 1 \mid S_{t-1} = 1) = p \\
P(S_t &= 0 \mid S_{t-1} = 0) = q
\end{aligned}
$$

其中，$\varphi(L) = (1 - \varphi_1 L - \cdots - \varphi_4 L^4) = 0$ 的根落在单位圆外面，y_t 是第 t 期实际 GDP 或 GNP 的对数值，Δy_t 可用来表示第 t 期实际 GDP 或 GNP 的增长率。

Kim，Nelson（1999）利用 Hamilton（1989）中的样本数据，即从 1952 年第二季度到 1984 年第四季度美国季度 GNP 数据，采用 EM 算法估计模型式 10.5。从得到的参数估计和状态序列估计来看，Kim，Nelson（1999）和 Hamilton（1989）的结

果非常接近。Kim，Nelson（1999）的估计结果在表10－1中给出。图10－3到图10－5分别描述了衰退时期的滤波概率 $P(S_t = 0 \mid y_1^t)$、平滑概率 $P(S_t = 0 \mid y_1^T)$ 和一步预测概率 $P(S_t = 0 \mid y_1^{t-1})$。MS－AR（4）模型的计算结果与美国国家经济研究局的周期高度吻合。

表10－1 Hamilton（1989）MS－AR（4）的参数估计

参数	估计值	标准差
p	0.9008	0.0443
q	0.7606	0.1206
φ_1	0.0898	0.1981
φ_2	－0.0186	0.2082
φ_3	－0.1743	0.1381
φ_4	－0.0839	0.1248
σ	0.7962	0.0858
μ_0	－0.2132	0.2613
μ_1	1.1283	0.1596

数据来源：Kim，Nelson（1999）。

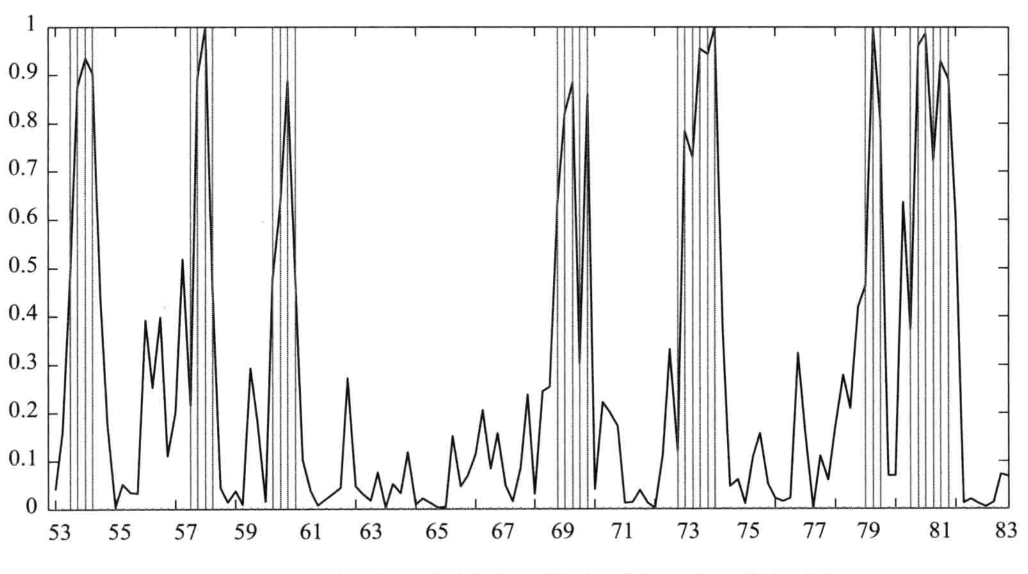


图10－3 衰退时期的滤波概率（GDP：1952：Ⅱ－1984：Ⅳ）

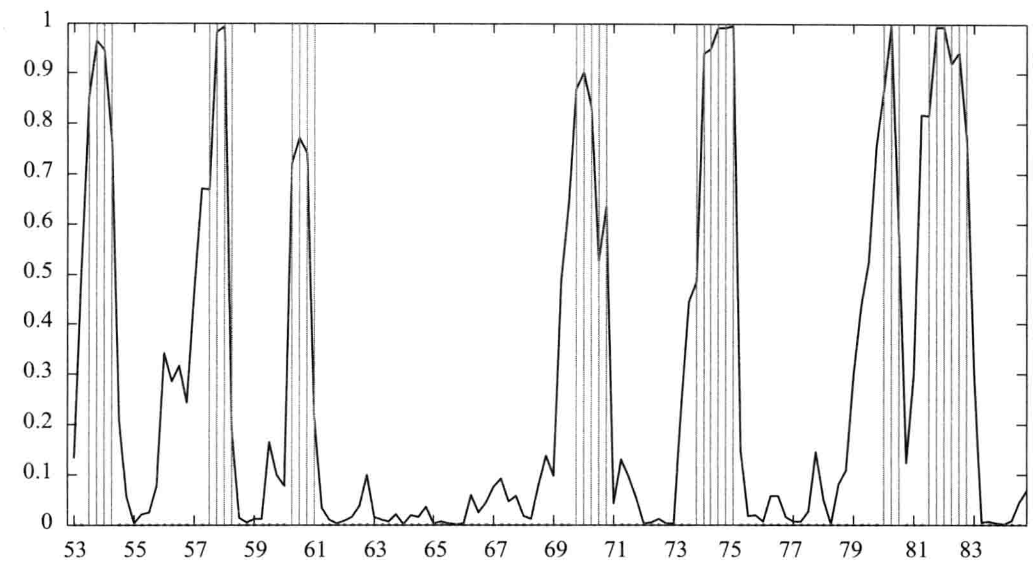


图 10－4　衰退时期的平滑概率（GDP：1952：Ⅱ－1984：Ⅳ）

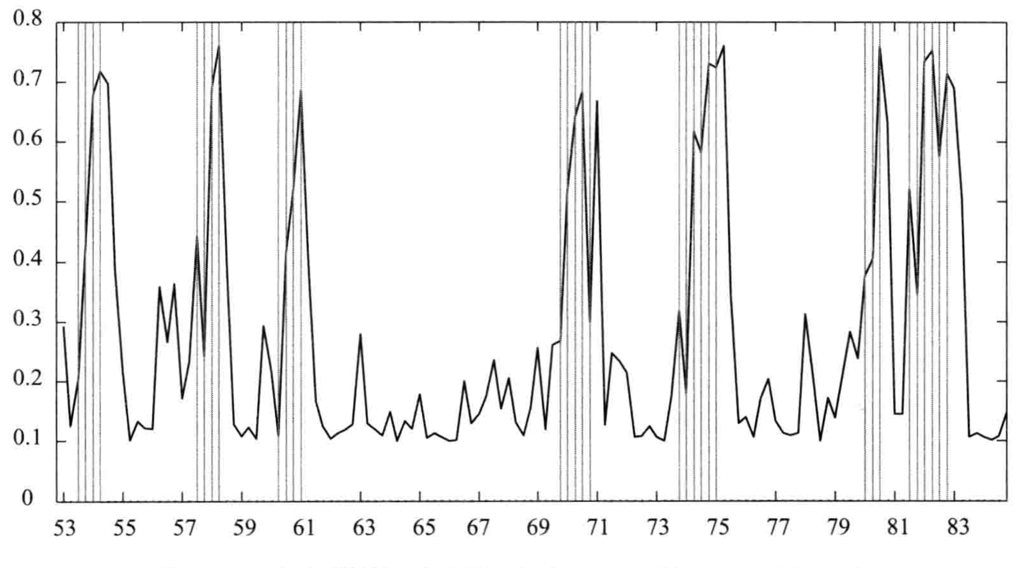


图 10－5　衰退时期的一步预测概率（GDP：1952：Ⅱ－1984：Ⅳ）

第三节 Kim，Nelson（1999）加入虚拟变量的 MS - AR (4) 模型

经济经历着一次又一次的衰退和膨胀，但值得注意的是，这些衰退阶段之间或膨胀阶段之间也是存在区别的。比如，上一次牛市阶段能够达到日均回报率不一定成为下一次牛市中指数的日均回报率，上一次经济膨胀中经济的年均增长率不一定会在下一个膨胀阶段中得到同样的保持。从分类方法的角度看，样本不仅存在组间差别，同样存在组内差别。当组内差别较为显著的时候，就需要对模型设置进行一定的纠正，从而能更准确地对状态变量进行估计。

Kim，Nelson（1999）从这样的思考角度发现：如果在样本中再加入一些年份的数据（1952：II－1995：III），模型难以提供合理的估计参数，因而就难以推断出合理的衰退或繁荣的概率。可能的主要原因是模型并没有考虑 20 世纪 90 年代美国的劳动生产率下降的问题；另一个原因可能是，即使美国这一时期不存在劳动生产率下降的问题，货币政策对经济稳定有着越来越重要的作用。考虑到上述因素，Kim，Nelson（1999）将模型的均值方程修改为：

$$\mu_{s_t} = (\mu_0 + \mu_0^* \cdot D_t)(1 - S_t) + (\mu_1 + \mu_1^* \cdot D_t)S_t \tag{10.6}$$

其中，D_t 是一个虚拟变量，样本为 1983：I－1995：III 时取 1，为早期样本时取 0。虚拟变量的引入潜在地控制了在繁荣或衰退期间平均增长率的变化。

表 10－2 带虚拟变量的 MS－AR（4）模型参数估计

	膨胀和衰退时均值都有变化		仅膨胀阶段均值有变化	
参数	估计值	标准差	估计值	标准差
p	0.9113	0.0363	0.9187	0.0309
q	0.7658	0.0357	0.7668	0.0863
φ_1	0.0496	0.1347	0.0477	0.1117

续表

	膨胀和衰退时均值都有变化		仅膨胀阶段均值有变化	
参数	估计值	标准差	估计值	标准差
φ_2	-0.0495	0.1295	-0.0422	0.1103
φ_3	-0.2112	0.1129	-0.2095	0.1008
φ_4	-0.0953	0.1140	-0.0984	0.0970
σ^2	0.6902	0.0505	0.6939	0.0474
μ_0	-0.2996	0.1392	-0.2328	0.1895
μ_1	1.1479	0.0768	1.1510	0.0776
μ_0^*	0.4516	0.3209	—	—
μ_1^*	-0.3346	0.1240	-0.3699	0.1244
极大似然值	-212.17		-212.99	

数据来源：Kim，Nelson（1999）。

表 10-2 的第 2 列和第 4 列给出了上述模型的估计参数。图 10-9 和图 10-10 给出了衰退时期的过滤概率和平滑概率。这些概率和美国国家经济研究局的参考周期吻合度很高。

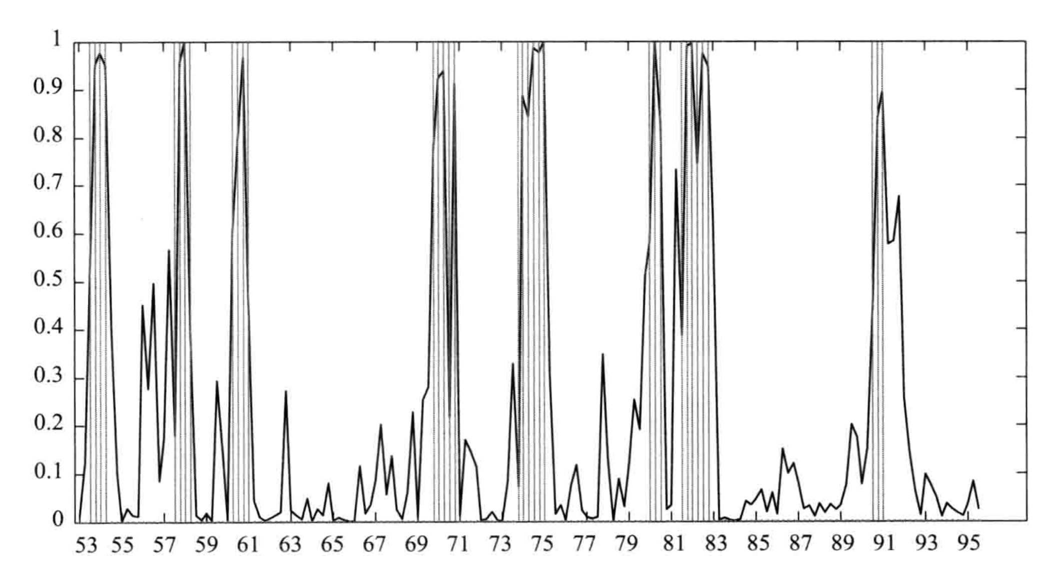


图 10-6　衰退时期的滤波概率（GDP：1952：Ⅱ-1995：Ⅲ）

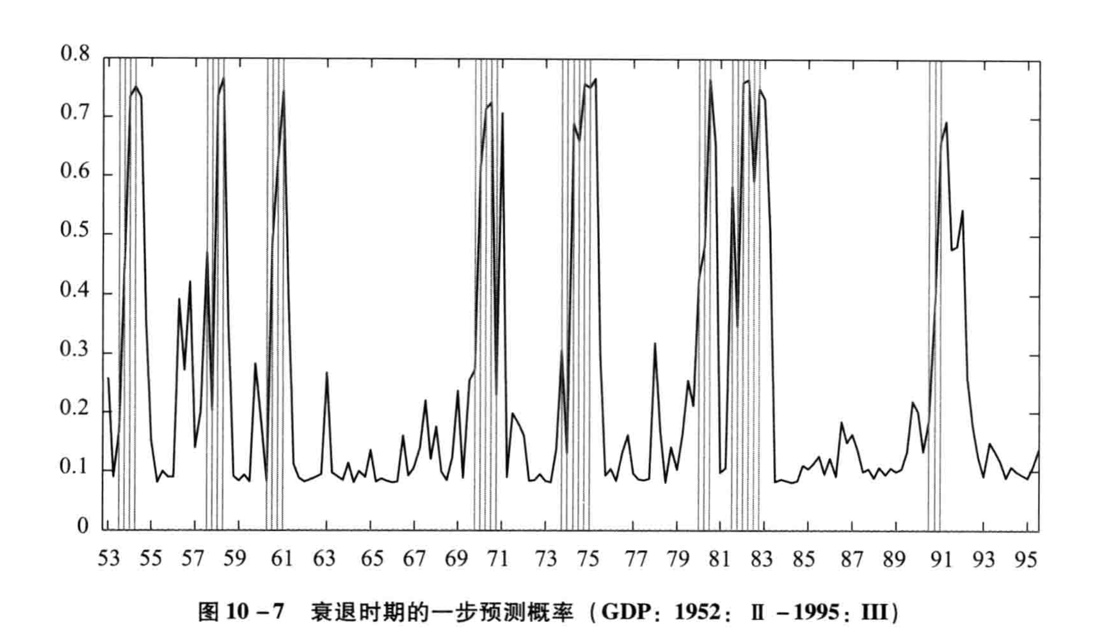


图 10－7 衰退时期的一步预测概率（GDP：1952：Ⅱ－1995：Ⅲ）

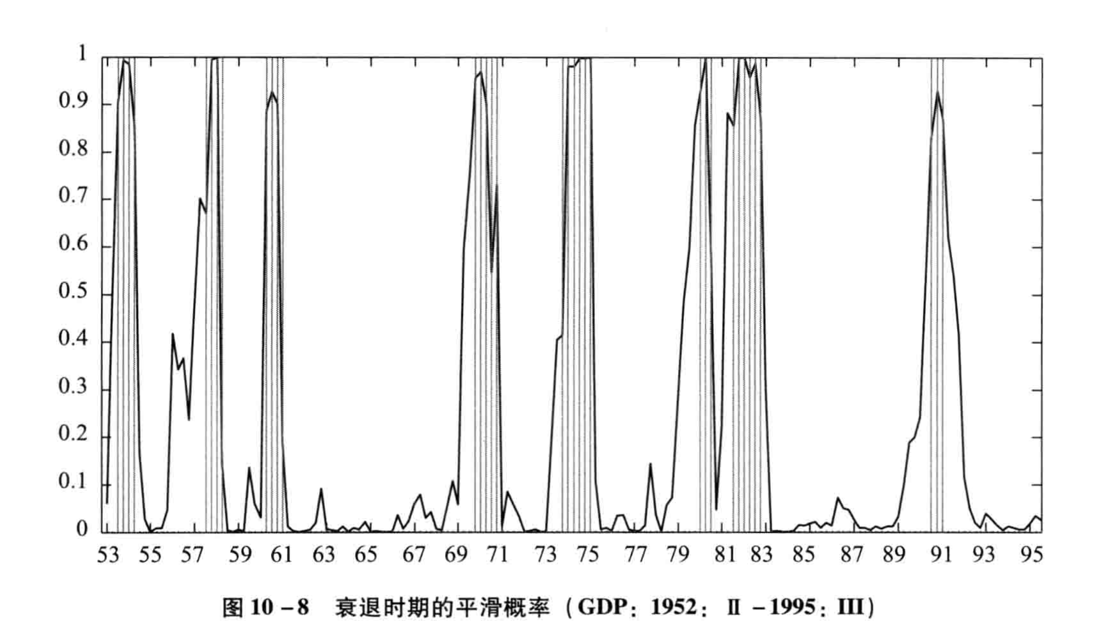
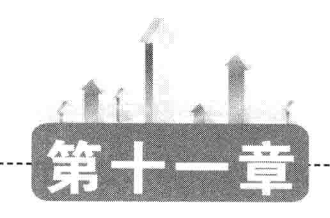


图 10－8 衰退时期的平滑概率（GDP：1952：Ⅱ－1995：Ⅲ）

该模型不仅仅对马尔科夫状态转换模型在宏观经济分析中起到了完善的作用，同时也说明，过往经济膨胀或萧条的特征虽然可以重现，但是不能机械地认为历史

能够完全不变地被复制。过往各期状态中因变量的分布可以用来预测本期因变量的分布，但也会有所差别。在利用马尔科夫状态转换模型进行状态分析时，模型设置会影响状态的识别，而状态的识别又会影响模型的设置。从这个角度来看，如果状态识别不能令人满意，那么则需要从数据特征的认识等方面再进行调整，这也符合计量模型建立的一般思路。



第十一章 HMM和SWARCH模型在股市中的应用

大自然中万物繁衍生息、新旧更替，都遵循一定的规律。这些规律或为人知或不为人知，但其是否发挥作用以及作用的大小都不会因为是否被人类知晓而发生任何变化。在经济、金融领域也是如此。金融市场与宏观经济的一个相似之处在于，两者都会发生阶段性的经济形势变化，这种变化在宏观经济中体现为经济周期运动，在金融市场尤其是股票市场中体现为牛市和熊市的交替。也就是说，它们在一定时期都会体现出特定阶段的规律。

认识和判断经济和金融市场运行的大趋势具有重要的指导意义。只有顺势而为才能在更大程度上规避风险，或理智而聪明地承担风险。对金融市场来说，虽然单个金融产品有可能保持自己独特的运行方式，不一定与市场整体表现有较强的相关性，但是却容易受到系统性风险的影响。而对于大多数金融产品来讲，很难在市场不景气的时候独善其身，也容易在市场繁荣的时候表现更为强劲。同时，有些金融产品与经济周期或市场趋势保持着稳定的正向或反向的关系，利用对市场行情的预测可以有效地提高投资回报水平。即便单个金融产品与市场整体表现关系不大，但自身运行也可能具有一定的阶段性，如果能准确判断出特定时期或未来的阶段特征，也能够带来较高的回报水平。基于这种对金融产品运动趋势和周期的判断能力在投资领域被称为择时能力。

金融市场不会永远都保持不变，不允许线性的思维方式，不会因为投资人的主

观期望而永远向上或向下运动，且必定有所反复。这个重要特征可以用马尔科夫模型来模拟：假定金融市场的运动可以分为若干个形态，而每个形态之存在相互转化、相互联系的动态关系，即服从马尔科夫过程。困难在于有多少个形态存在、每个形态在何时存在、当前及接下来的形态如何、每一个形态的特征如何，等等，这些都无法从金融市场中直接观测得到。形态数量确定和每一个形态特征需要解决模型设置和估计问题；每个形态何时存在以及当前和接下来形态如何需要解决学习问题。这两个问题其实就是我们前面提到的 EM 解法的两个基本步骤。

根据有效市场理论，通过量化分析的方法得到金融市场阶段划分、反转点、收益风险阶段特征等信息之后，发现了原来不为大多数投资者所知晓的信息，这相当于增加了市场信息量。如果有效的量化投资技术被少部分投资者所掌握，那么这部分投资者会更容易获得持续的超额收益，从而证伪市场有效性理论；反之，如果有效的量化投资技术被更大多数投资者所掌握，那么这种获利空间就会减少，而更多地体现为市场有效性的提高。所以，量化投资手段的应用在短期可以为投资者带来较为丰厚的投资回报，在长期可以提高市场有效性①。

本章利用马尔科夫模型来研究金融市场指数。首先，介绍上证综指历史数据基本特征和所选取数据；然后，说明马尔科夫模型设置情况和估计方法；最后，针对得到的结论分析如何将马尔科夫模型应用于市场行情的判断。这种研究方法可以扩展到对行业、板块或个股的研究。

① 随机游走假说等市场有效性判断方法将过去收益率或指数数据作为以往信息的代理变量，通过考察自相关的形式来判断市场是否有效。这种研究方法并未将 HMM 以及 MS 的情况考虑进来。可能的情况是：观测值序列并不直接相关，而是通过潜在的不可观测的状态变量的相关性来彼此联系，从而发挥历史信息的预测作用。

第一节 股指收益率与 HMM

收益率特征是投资者对股票市场关注的重点，如何通过收益率将股票市场运行划分为牛市和熊市阶段有助于投资者做出正确键的投资判断。这里利用 HMM 并选取 2000 年 1 月至 2013 年 12 月上证综指月度数据进行分析。在分析之前，通过对数差分方法将指数历史水平序列转化为收益率序列。

模型设置考虑到以下因果关系：假设在给定 t 期的状态变量之后，这 t 期因变量的分布不再取决于其滞后值或滞后期的状态变量。用图形来表示其中的因果关系如下：

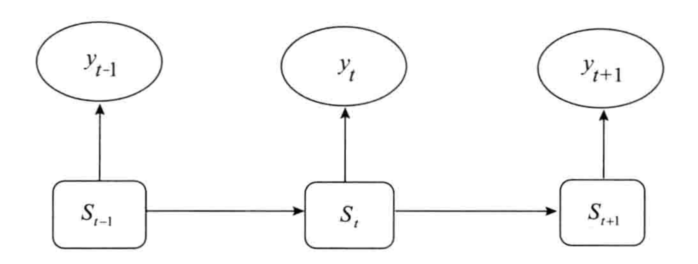


图 11－1　序列不相关马尔科夫均值转换模型因果关系示意图

这意味着观测值序列，也就是指数收益率序列 y_t 的分布仅仅取决于 S_t，而与 $y_{t-1},y_{t-2},\cdots,y_1$ 等历史收益率都无关。这是一种简化的做法，目的是方便分析，重点强调问题分析的来龙去脉，在实践中，可以考虑序列相关的情况，在某些情况下这也是非常有必要的。根据这样的因果关系，得到下面形式设置的模型：

$$y_t = \mu_{S_t} + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N(0,\sigma^2) \tag{11.1}$$

$$\mu_{S_t} = (1 - S_t) \cdot \mu_0 + S_t \cdot \mu_1 \tag{11.2}$$

$$P(S_t = 0 \mid S_{t-1} = 0) = p \tag{11.3}$$

$$P(S_t = 1 \mid S_{t-1} = 1) = q \tag{11.4}$$

利用极大似然估计法估计模型中的参数。从状态转换的平均值估计值来看：μ_0

= - 0.0182，p 统计量为 0.00；μ_1 = 0.0913，p 统计量为 0.07。两者均显著异于零。从均值来看，状态 0 对应的是股市中的熊市，而状态 1 对应着股市中的牛市。方差 σ^2 不随状态变量变化，其估计值为 0.005228，p 统计量为 0.00，同样显著异于零。状态转换矩阵为：

$$\begin{bmatrix} 0.98 & 0.02 \\ 0.09 & 0.91 \end{bmatrix}$$

所以，熊市的自我转换概率更高，为 0.98；牛市的自我转换概率较低，为 0.91。

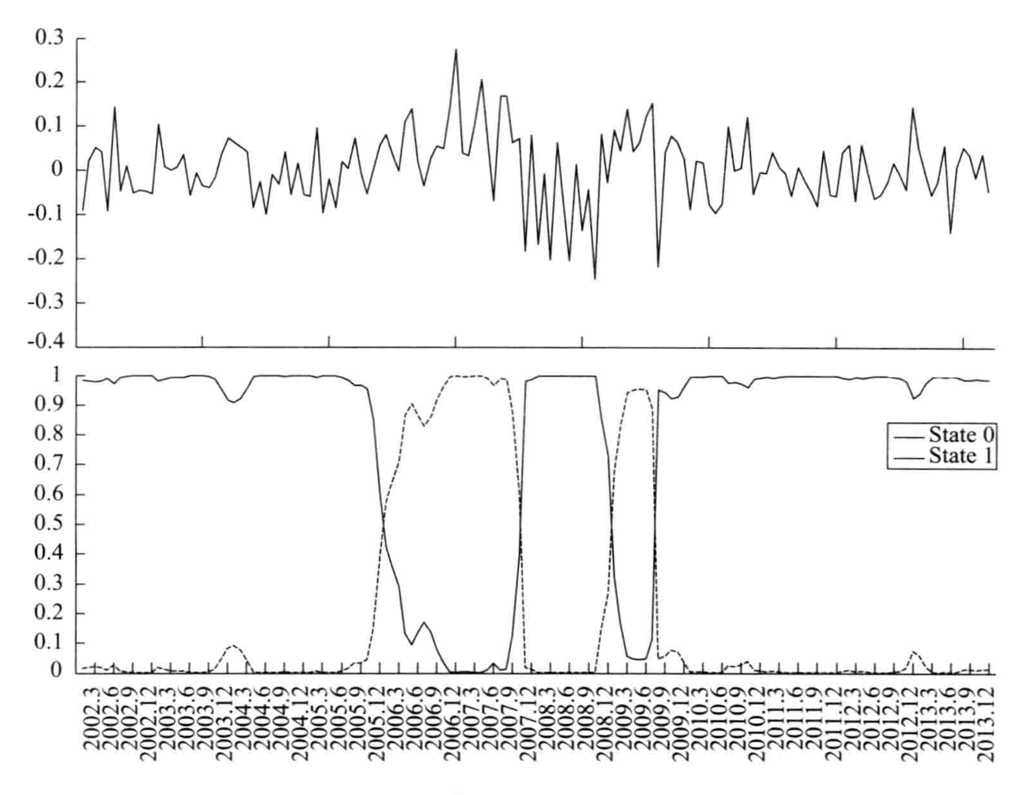


图 11－2　上证综指月度收益率（2000－2013 年）及平滑概率

从上面图 11－2 平滑概率可知：状态 0 用实线表示，对应着熊市的概率；状态 1 用虚线表示，对应着牛市的概率。在 2005 年 12 月至 2007 年 10 月期间，以及 2009 年 1 月至 2009 年 7 月期间，牛市的概率大于熊市的概率，市场呈现上升势头；在其余期间市场长期呈现熊市特征。从理论上来看，牛市的平均持续期为 10.63 个

月，熊市持续期为43.45个月。但不幸的是，虽然市场已经度过了四年多（大致53个月）的熊市时期，但截至2013年末，仍然处于市场较弱的熊市阶段，并没有明显迹象表明市场具有进入牛市趋势。基于上述判断，可以得到下面策略选择：

1. 基于平滑概率的操作策略

从图11－2中可以明显看到上证综指变化的阶段特征，基于这样的判断，可以转换投资风格和风险暴露程度，以便在市场弱势时控制风险，在市场强势时博取更高的收益。根据资本资产定价模型，承担的市场风险与所获得的投资回报是正相关的。那么，可以在牛市时选择承担更多的系统性风险，从而获得更高的投资回报；在熊市时选择承担较少的系统性风险，从而避免潜在的投资损失。或者遵守以下策略：在牛市时期选择贝塔值较大的股票，在熊市时选择贝塔值较小的股票；在牛市时采取跟踪股指的被动投资策略，在熊市时发挥择股能力寻求稀缺的投资机会。总之，将市场分为熊市和牛市两个阶段，在不同阶段采取不同的投资策略，相比缺少市场划分或划分可靠性不高的情况，更有利于提高投资回报。

2. 滤波概率与平滑概率结合的操作策略

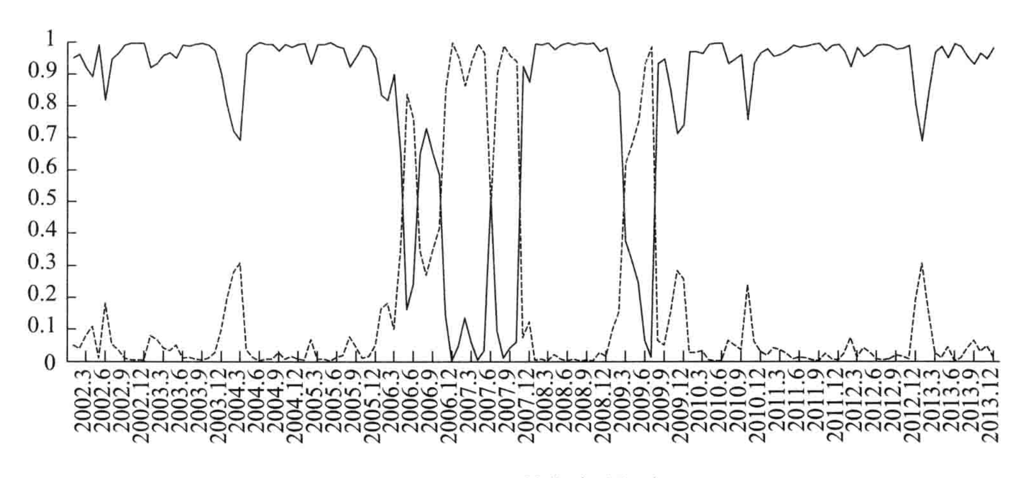


图11－3　上证综指滤波概率

图11－3为滤波概率，利用的是各时点当期的信息，而不是整个样本期的信息。从图中可以发现，平滑概率和滤波概率两者大致相同，但也有一些差别。比如，滤

波概率的平滑程度较低，局部可能出现概率的突然增加。这意味着基于滤波概率的市场操作策略会更为频繁地改变方向，如果判断有误，那么频繁的操作将加大损失的程度。而平滑概率则比较一致，不会出现操作方向的突变。其实，这种特点并不是滤波概率的缺点，因为在局部出现的概率突变可能意味着较为短期的投资时机，也就是牛市中的小熊市或熊市中的小牛市。这是符合股市变化一般规律的，如果抓住这些零散的投资时机，那么也能获得较高的投资回报。

投资者们普遍感觉，在当前的这个时期内很难获得较高的投资收益，这与我国股票市场所处的阶段是分不开的。只有具备较强的择股能力才能在困境中寻找到合适的投资标的，从而取得超越市场的投资回报。下面将马尔科夫状态转换模型应用到对个股的研究中。

第二节　股指波动性与 SWARCH 模型

Kim，Nelson 和 Startz（1998）将三个状态的马尔科夫方差转换模型运用到 1926 年 1 月至 1986 年 12 月期间股票月收益数据中，以此来处理数据异方差问题。我们现在关注的是如何利用三状态马尔科夫转换模型来对股票收益做出可靠的预测和估计。

Fama（1963）和 Mandlebrot（1963）指出，股票收益服从具有一定偏度和较大峰度的非正态分布。Turner，Startz 和 Nelson（1989）也同样指出，股票市场收益分布具有典型的高峰、厚尾、条件异方差等特征。因此，Engle（1982）和 Bollerslev（1986）将 ARCH 模型应用在股票收益率的条件异方差研究中。另外有关股票收益分布的建模方法上也不仅局限于一个或一种分布，而是假定收益率服从由多个分布组成的混合分布，从而解决了单一、非时变方差模型存在的问题，我们在第二部分对此方法进行过详细的讨论。

利用马尔科夫机制转换模型同样可以较好地拟合股票市场收益率序列的方差。在较早期的研究中，Hamilton 和 Susmel（1994）提出了 SWARCH（Switching

ARCH）模型。在这个模型中，他们允许 ARCH 模型的系数来自几个不同状态下的参数组合，每一个状态对应一个参数组合。同前面的分析一样，这种指示关系通过状态变量来实现。Hamilton 和 Susmel（1994）将 SWARCH 模型应用在股票每周的收益率数据上，他们发现 ARCH 的影响在一个月后几乎完全消失。Hamilton 和 Susmel（1994）SWARCH 模型的简化形式如下：

$$y_t = \sigma_{s_t} v_t \tag{11.5}$$

$$v_t = h_t \varepsilon_t \text{ , } \varepsilon_t \sim i.i.d.t \tag{11.6}$$

$$h_t = \alpha_0 + \alpha_1 v_{t-1}^2 + \alpha_2 v_{t-2}^2 + \beta d_{t-1} v_{t-1}^2 \tag{11.7}$$

其中 σ_{s_t} 是马尔可夫转换方差，d_{t-1} 是用来表示杠杆效应的虚拟变量。由于肥尾分布，t 分布能够更好地描述股票价格上升慢，下降快。Hamilton 和 Susmel 的估计结果显示 $\hat{\lambda} = \hat{\alpha}_1 + \hat{\alpha}_2 = 0.48$ 。这里，$\hat{\lambda}^4 = 0.05$ ，这说明受 μ_t 或 ARCH 影响产生的波动效应在一个月后几乎完全消失，所以，在为月度股票收益率建模的时候，并不需要考虑 ARCH 效用。

Kim，Nelson 和 Startz（1998）考虑了如下三种状态下的股票周收益马尔可夫转换模型：

$$y_t \sim N(0, \sigma_t^2) \tag{11.8}$$

$$\sigma_t^2 = \sigma_1^2 S_{1t} + \sigma_2^2 S_{2t} + \sigma_3^2 S_{3t} \tag{11.9}$$

如果 $S_t = k \ (k = 1,2,3)$，则 $S_{kt} = 1$ ；否则 $S_{kt} = 0$ 。

$$P(S_t = j \mid S_{t-1} = i) = p_{ij} \quad i,j = 1,2,3 \tag{11.10}$$

$$\sum_{j=1}^{3} p_{ij} = 1$$

$$\sigma_1^2 < \sigma_2^2 < \sigma_3^2 \tag{11.11}$$

其中，y_t 是去均值后的月股票收益率，S_t 是不可观测的状态变量，且由具有转换概率的一阶马尔科夫过程演变而来。式 11.11 是模型识别的必要条件，表 11－2 给出了相关的参数估计和标准误差。

表 11－2　异方差下股票收益的三状态马尔可夫转换模型的极大似然估计

参数	估计值	方差
p_{11}	0.9736	0.0177
p_{12}	0.0264	0.0177
p_{21}	0.0197	0.0120
p_{22}	0.9686	0.0142
p_{31}	0.0028	0.0193
p_{32}	0.0460	0.0376
σ_1^2	0.0012	0.0002
σ_2^2	0.0040	0.0005
σ_3^2	0.0310	0.0057
似然值	－1001.90	

Kim 等人检验了三状态马尔可夫转换方差，结果并没有发现 ARCH 效应。这与 Hamilton 和 Susmel 等人的结果一致。此外，标准化收益率序列的峰值并不明显，Jarque－Bera 正态检验的 p 值为 0.073，也就是说，在 5% 的置信水平上并不能拒绝标准化收益率是正态分布的原假设。这些结果表明，三状态的马尔科夫转换方差模型为 1926 年 1 月至 1986 年 12 月股票月收益率异方差性提供了一个可信的解释。

下面利用 SWARCH 模型来探讨我国股市波动性问题，选取 1991 年 1 月至 2013 年 12 月上证综指的周度历史收益率数据进行分析。

利用模型选取准则确定状态个数为 2。而且从三状态的机制转换 ARCH 模型可以发现，仅有两个状态明显主导了整个市场波动情况，第三个状态可以忽略不计。因此，波动状态个数确定为两个。同样利用模型选取准则最终确定采取 SWARCH 模型来进行估计①，其中条件方差为一阶自相关过程。利用极大似然算法求得模型参数和波动性状态。我们期望达到两个目的：

首先，波动性参数是资本市场定价模型和投资决策的基础，也是绩效考核的关键，对波动性更加准确的估计有利于有效控制投资风险，更加聪明地承担必要的市场风险，并科学地考核投资业绩。利用 SWARCH 模型和表 11－2 中的参数估计，可

① 在这里，观测值序列一阶自相关或无自相关对研究结论影响不大。

以形成上证综指波动性的有效预期。

表 11－3 上证综指收益率 SWARCH 模型参数估计

参数	估计值	方差
p_{11}	0.9973	0.0021
p_{22}	0.9922	0.0786
σ_1^2	0.0540	0.0032
σ_2^2	0.0757	0.0005
似然值	－3463.2813	

其次，可以通过市场波动性的阶段划分来确定市场阶段性变化。从图 11－4 中可以看出，我国股市大致经历了两个波动性较大的时期：1997 年之前，以及 2006 年底至 2009 年底。第一个时期的高波动性是因为我国股市在 1996 年 12 月 26 日才开始实行涨停板制度，所以股指相对来说保持了较高的波动性；第二个时期的高波动性对应着股指水平冲击历史高位的时期，股指从 2000 点以下一直冲到 6124 点，但是又在接下来的一年深度下跌到 2000 点以下，之后的一年保持增长趋势，指数反弹到 3500 左右。但之后整体呈现下跌趋势，市场波动性也有大幅度下降。结合图 12－2，从收益率和波动性综合角度看，市场牛市和熊市相互交替的阶段往往伴随着市场波动性的放大。这也为寻找市场反转时机提供了一定的依据。

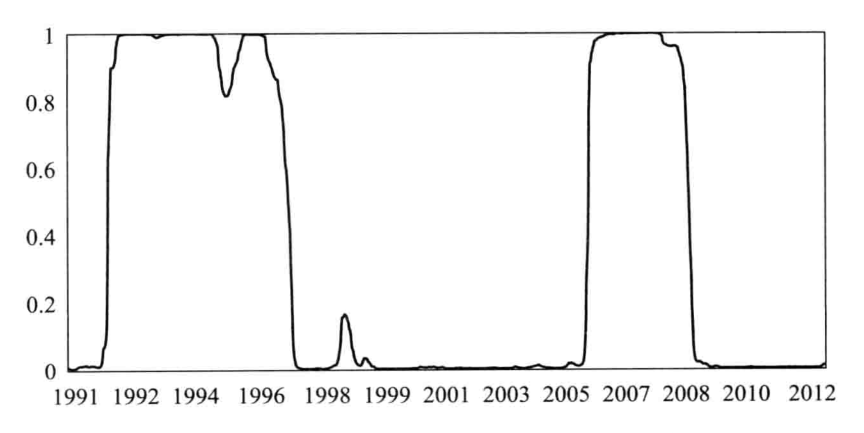


图 11－4 上证综指处于高波动时期的平滑概率

参考文献

1. Abramowitz, M. , Stegun, I. A. , Danos, M. and Rafelski, J. (eds.) . Pocketbook of Mathematical Functions. Abridged Edition of Handbook of Mathematical Functions. Verlag Harri Deutsch, Thun and Frankfurt am Main, 1984.

2. Aitchison, J. The statistical analysis of compositional data [J]. J. Roy. Statist. 1982, Soc. B 44: 139 – 177.

3. Albert, P. S. A two – state Markov mixture model for a time series of epileptic seizure counts [J]. Biometrics 47, 1991: 1371 – 1381.

4. Alexander, C. Market Risk Analysis: Practical Financial Econometrics [M]. Wiley, 2008.

5. Altman, R. MacKay and Petkau, J. A. Application of hidden Markov models to multiple sclerosis lesion count data [J]. Statist. Med. 24, 2005: 2335 – 2344.

6. Altman, R. MacKay. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting [J]. J. Amer. Statist. Assoc. 2007 (102): 201 – 210.

7. Andrews, D. W. K. Tests for Parameter Instability and Structural Change with Unknown Change Point [J]. Econometrica. 1993 (62): 1383 – 1414.

8. Aston, J. A. D. and Martin, D. E. K. Distributions associated with general runs and patterns in hidden Markov models [J]. Ann. Appl. Statist. 2007 (1): 585 – 611.

9. Azzalini, A. and Bowman, A. W. A look at some data on the Old Faithful geyser [J]. Appl. Statist. 1990 (39): 357 – 365.

10. Barton Browne, L. Physiologically induced changes in resourceoriented behaviour [J]. Ann. Rev.

Entomology. 1993 (38): 1 - 25.

11. Baum, L. E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. //O. Shisha (ed.). Proc. Third Symposium on Inequalities. New York: Academic Press, 1972: 1 - 8.

12. Baum, L. E., Petrie, T., Soules, G. and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains [J]. Ann. Math. Statist. 1970 (41): 164 - 171.

13. Bellman, R. Introduction to Matrix Analysis [M]. New York: McGraw - Hill, 1960.

14. Berchtold, A. and Raftery, A. E. The mixture transition distribution model for high - order Markov chains and non - Gaussian time series. Statist [J]. Sci. 2002 (17): 328 - 356.

15. Berchtold, A. Estimation in the mixture transition distribution model [J]. J. Time Series Anal. 2001 (22): 379 - 397.

16. Berchtold, A. The double chain Markov model [J]. Commun. Stat. Theory Meth. 1999 (28): 2569 - 2589.

17. Bisgaard, S. and Travis, L. E. Existence and uniqueness of the solution of the likelihood equations for binary Markov chains [J]. Statist. Prob. Letters. 1991 (12): 29 - 35.

18. Bishop, C. M. Pattern Recognition and Machine Learning [M]. New York: Springer, 2006.

19. Bollerslev, Tim. Generalized Autoregressive Conditional Heteroskedasticity [J]. Journal of Econometrics. 1986 (31): 307 - 317.

20. Bollerslev, Tim, Ray Y. Chou, and Kenneth F. Kroner. ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence [J]. Journal of Econometrics. 1992 (52): 5 - 59.

21. Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. Time Series Analysis, Forecasting and Control, third edition [M]. Englewood Cliffs: Prentice Hall, 1994.

22. Boys, R. J. and Henderson, D. A. A Bayesian approach to DNA sequence segmentation (with discussion) [J]. Biometrics. 2004 (60): 573 - 588.

23. Brockwell, A. E. Universal residuals: A multivariate transformation [J]. Statist. Prob. Letters. 2007 (77): 1473 - 1478.

24. Brooks, C. Introduction to Econometrics [M]. Cambridge University Press, 2002.

25. Brown, R. L., J. Durbin, and J. M. Eavns. Techniques for Testing the Constancy of Regression Relationships over Time [J]. Journal of the Royal Statistical Society. 1975, B37: 149 - 192.

26. Bulla, J. and Berzel, A. Computational issues in parameter estimation for stationary hidden Markov models [J]. Computat. Statist. 2008 (23): 1 - 18.

27. Bulla, J. and Bulla, I. Stylized facts of financial time series and hidden semi - Markov models. Computat [J]. Statist. & Data Analysis. 2007 (51): 2192 - 2209.

28. Calvet L. and Fisher, A. J. How to forecast long - run volatility: regime switching and the estimation of multifractal processes [J]. J. Financial Econometrics. 2004 (2): 49 - 83.

29. Calvet, L. and Fisher, A. J. Forecasting multifractal volatility [J]. J. Econometrics. 2001 (105): 27 - 58.

30. Cappé, O., Moulines, E. and Rydén, T. Inference in Hidden Markov Models [M]. New York: Springer, 2005.

31. Celeux, G., Hurn, M. and Robert, C. P. Computational and inferential difficulties with mixture posterior distributions [J]. J. Amer. Statist. Assoc. 2000 (95): 957 - 970.

32. Chib, S. Calculating posterior distributions and modal estimates in Markov mixture models [J]. J. Econometrics. 1996 (75): 79 - 97.

33. Chopin, N. Inference and model choice for sequentially ordered hidden Markov models [J]. J. Roy. Statist. 2007, Soc. B 69: 269 - 284.

34. Chow, G. Tests of the Equality berween Two Sets of Coefficients in Two Linear Regressions [J]. Econometrica. 1960 (28): 561 - 605.

35. Congdon, P. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities [J]. Computat. Statist. & Data Analysis. 2006 (50): 346 - 357.

36. Cook, R. D. and Weisberg, S. Residuals and Influence in Regression [M]. London: Chapman & Hall, 1982.

37. Cosslett, S. R. and Lee, L. - F. Serial correlation in latent discrete variable models [J]. J. Econometrics. 1985 (27): 79 - 97.

38. Cox, D. R. and Snell, E. J. A general definition of residuals (with discussion) [J]. J. Roy. Statist. 1968, Soc. B 30: 248 – 275.

39. Cox, D. R. Role of models in statistical analysis [J]. Statist. Sci. 1990 (5): 169 – 174.

40. Cox, D. R. Statistical analysis of time series: some recent developments [J]. Scand. J. Statist. 1981 (8): 93 – 115.

41. Cramer, J. S. Econometric Applications of Maximum Likelihood Methods [M]. Cambridge: Cambridge University Press, 1986.

42. Davidson, Russell, and James G. MacKinnon. Estimation and Inference in Econometrics [M]. Oxford, UK: Oxford University Press, 1993.

43. Davison, A. C. Statistical Models [M]. Cambridge: Cambridge University Press, 2003.

44. Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion) [J]. J. Roy. Statist. 1977, Soc. B 39: 1 – 38.

45. Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society. 1977, B39: 1 – 38.

46. Dewsbury, D. A. On the problems studied in ethology, comparative psychology, and animal behaviour [J]. Ethology. 1992 (92): 89 – 107.

47. Diebold, Francis X., Joon – Haeng Lee, and Gretchen C. Weinbach. Regime Switching with Time – Varying Transition Probabilities. // C. Hargreaves (ed.). Nonstationary Time Series Analysis and Cointegration. Oxford: Oxford University Press, 1994: 283 – 302.

48. Diggle, P. J. Contribution to the discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods [J]. J. Roy. Statist. 1993, Soc. B 55: 67 – 68.

49. Draper, D. Contribution to the discussion of Raftery et al. 2007: 36 – 37.

50. Dunn, P. K. and Smyth, G. K. Randomized quantile residuals [J]. J. Comp. Graphical Statist. 1996 (5): 236 – 244.

51. Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids [M]. Cambridge: Cambridge University Press, 1998.

52. Durland, J. Michael, and Thomas H. McCurdy. Duration – Dependent Transitions in a Markov Model of U. S. GNP Growth [J]. Journal of Business and Economic Statistics. 1994 (12): 279 –288.

53. Efron, B. and Tibshirani, R. J. An Introduction to the Bootstrap [M]. New: York Chapman & Hall, 1993.

54. Engel, Charle, and James Hamilton. Long Swings in the Dollar: Are They in the Data and Do Markets Know It? [J] American Economic Review. 1990, 80 (4): 689 –713.

55. Engle, Robert F. Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation [J]. Econometrica. 1982 (50): 987 –1007.

56. Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. IEEE Trans. Inform. Th. 48: 1518 –1569.

57. Fama, Eugene F. Mandlebrot and the Stable Paretian Hypothesis [J]. Journal of Business. 1963 (4): 420 –429.

58. Fama, Eugene F. Short – Term Interest Rates as predictors of Inflation [J]. American Economic Review. 1975 (65): 269 –282.

59. Farley, J. U. , and M. J. Hinich. A Test for a Shifting Slope Coefficient in a Linear Model [J]. Journal of the American Statistical Association. 1970 (65): 1320 –1329.

60. Feller, W. An Introduction to Probability Theory and Its Applications [M], Volume 1, third edition. New York: Wiley, 1968.

61. Filardo Andrew J. Business Cycle Phases and Their Transitional Dynamics [J]. Jurnal of Business and Economic Statistics. 1994 (12): 29 –308.

62. Fisher, N. I. and Lee, A. J. A correlation coefficient for circular data [J]. Biometrika. 1983 (70): 327 –332.

63. Fisher, N. I. and Lee, A. J. Time series analysis of circular data [J]. J. Roy. Statist. 1994, Soc. B 56: 327 –339.

64. Fisher, N. I. The Analysis of Circular Data [M]. Cambridge: Cambridge University Press, 1993.

65. Forney, G. D. The Viterbi algorithm. Proc. IEEE. 1973 (61): 268 –278.

66. Franke J, Seligmann T. Conditional maximum likelihood estimates for INAR (1) processes and their application to modelling epileptic seizure counts. // T. Subba Rao (ed.) Developments in Time Series Analysis. London: Chapman & Hall, 1993: 310 – 330.

67. Fredkin, D. R. and Rice, J. A. Bayesian restoration of single – channel patch clamp recordings [J]. Biometrics. 1992 (48): 427 – 448.

68. Fridman, M. and Harris, L. A maximum likelihood approach for non – Gaussian stochastic volatility models [J]. J. Bus. Econ. Statist. 1998 (16): 284 – 291.

69. Frühwirth – Schnatter, S. Finite Mixture and Markov Switching Models [M]. New York: Springer, 2006.

70. Garbade, K., and P. Wachtel. Time Variation in the Relationship between Inflation and Interest Rates [J]. Journal of Monetary Economics. 1978 (4): 755 – 765.

71. Garcia, Rene, and Pierre Perron. An Analysis of Real Interest under Regime Shift [J]. Review of Economics and Statistics. 1996 (78): 111 – 125.

72. Gill, P. E., Murray, W. and Wright, M. H. Practical Optimization [M]. London: Academic Press, 1981.

73. Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. User's Guide for NPSOL: a Fortran package for nonlinear programming [R]. Report SOL 86 – 2, Department of Operations Research, Stanford University, 1986.

74. Goldfeld, S. M., and R. E. Quandt. A Markov Model for Switching Regression [J]. Journal of Econometrics. 1973 (1): 3 – 16.

75. Gould, S. J. The Mismeasure of Man, revised and expanded edition [M]. London: Penguin Books, 1997.

76. Granger, C. W. J. Acronyms in time series analysis (ATSA) [J]. J. Time Series Anal. 1982 (3): 103 – 107.

77. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination [J]. Biometrika. 1995 (82): 711 – 732.

78. Grimmett, G. R. and Stirzaker, D. R. Probability and Random Processes, third edition [M].

Oxford: Oxford University Press, 2001.

79. Guttorp, P. Stochastic Modeling of Scientific Data [M]. London: Chapman & Hall, 1995.

80. Haines, L. M., Munoz, W. P. and van Gelderen, C. J. ARIMA modelling of birth data [J]. J. Appl. Statist. 1989 (16): 55-67.

81. Hamilton, J., D. Time Series Analysis [M]. Princeton University Press, 1994.

82. Hamilton, James D. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle [J]. Econometrica. 1989, 57 (2), 357-384.

83. Hamilton, James D. Analysis of Time Series Subject to Changes in Regime [J]. Journal of Econometrics. 1990 (45): 39-70.

84. Hamilton, James D. Estimation, Inference, and Forecasting of Time-Series Subject to Changes is Regime. // G. D. Maddala, C. R. Rao, and H. D. Vinod (ed.). Handbook of Statistics. New York: Elsevie Science Publishers B. V. 1993, Vol. 11: 231-259.

85. Hamilton, James D., and Raul Susmel. Autoregressive Conditional Heteroskedasticity and Changes in Regime [J]. Journal of Econometrics. 1994 (64): 307-333.

86. Haney, D. J. Methods for analyzing discrete-time, finite state Markov chains [D]. Ph. D. dissertation, Department of Statistics, Stanford University, 1993.

87. Hansen, Lars P. and Kenneth J. Singleton. Generalized Instrumenta Variables Estimation of Nonlinear Rational Expectations Models [J]. Econometrica, 1982 (50): 1269-1288.

88. Hansen, Lars P., and Kenneth J. Singleton. Stochastic Consumption, Risk Aversion and the Temporal Behavior of Asset Patterns [J]. Journal of Political Economy. 1983 (91): 249-265.

89. Harte, D. R package "Hidden Markov", version 1.2-5. http://www.statsresearch.co.nz, 27 July 2008.

90. Harvey, Andrew C. The Econometric Analysis of Time Series [M]. 2nd ed. MIT Press, 1990.

91. Harvey, Andrew C. Time Series Models [M]. Oxford: Philip Allan and Humanities Press, 1981.

92. Hasselblad, V. Estimation of finite mixtures of distributions from the exponential family [J]. J. Amer. Statist. Assoc. 1969 (64): 1459-1471.

93. Hastie, T., Tibshirani, R. J. and Friedman, J. The Elements of Statistical Learning: Data Min-

ing, Inference and Prediction [M]. New York: Springer, 2001.

94. Hastie, T. J. and Tibshirani, R. J. Generalized Additive Models [M]. London: Chapman & Hall, 1990.

95. Holzmann, H., Munk, A., Suster, M. L. and Zucchini, W. Hidden Markov models for circular and linear - circular time series [J]. Environ. Ecol. Stat. 2006 (13): 325 - 347.

96. Hopkins, A., Davies, P. and Dobson, C. Mathematical models of patterns of seizures: their use in the evaluation of drugs [J]. Arch. Neurol. 1985 (42): 463 - 467.

97. Hughes, J. P. A class of stochastic models for relating synoptic atmospheric patterns to local hydrologic phenomena [D]. Ph. D. dissertation, University of Washington, 1993.

98. Ihaka, R. and Gentleman, R. R: a language for data analysis and graphics [J]. J. Comp. Graphical Statist. 1996 (5): 299 - 314.

99. Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W. and Couto, E. Multistate Markov models for disease progression with classification error [J]. The Statistician. 2003 (52): 193 - 209.

100. Jacquier, E., Polson, N. G. and Rossi, P. E. Bayesian analysis of stochastic volatility models with fat - tails and correlated errors [J]. J. Econometrics. 2004 (122): 185 - 212.

101. Jammalamadaka, S. R. and Sarma, Y. R. (1988). A correlation coefficient for angular variables. // K. Matusita (ed.). Statistical Theory and Data Analysis II. New York: North Holland, 1988: 349 - 364.

102. Jammalamadaka, S. R. and SenGupta, A. Topics in Circular Statistics [M]. Singapore: World Scientific, 2001.

103. Jordan, M. I. Graphical models [J]. Statist. Science. 2004 (19): 140 - 155.

104. Juang, B. H. and Rabiner, L. R. Hidden Markov models for speech recognition [J]. Technometrics. 1991 (33): 251 - 272.

105. Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lutkepohl and T - C. Lee. Introduction to the Theory and Practice of Econometrics [M]. New York: John Wiley & Sons, 1982.

106. Kelly, F. P. Reversibility and Stochastic Networks [M]. Chichester: Wiley, 1979.

107. Kennedy, J. S. The New Anthropomorphism [M]. Cambridge: Cambridge University Press, 1992.

108. Kim, C., J., Nelson, C., R. State Space Model with Regime Switching: Classical and Gibbs – Sampling Approaches with Applications [M]. The MIT press, 1999.

109. Kim, S., Shephard, N. and Chib, S. Stochastic volatility: likelihood inference and comparison with ARCH models [J]. Rev. Econ. Studies. 1998 (65): 361 – 393.

110. Kim, Chang – Jin, and Charles R. Nelson. Business cycle Turning Points, A New Coincident Index, and Tests of Duration Dependence Based on A Dynamic Factor Model with Regime – Switching [J]. Review of Economics and Economic Statistics. 1998 (80): 188 – 201.

111. Kim, Chang – Jin, Charles R. Nelson, and Richard Startz. Testing for Mean Reversion in Heteroskedastic Data Based on Gibbs – Sampling – Augmented Randomization [J]. Journal of Empirical Finance. 1998 (5): 131 – 154.

112. Kim, Chang – Jin. Dynamic Linear Models with Markov – Switching [J]. Journal of Econometrics. 1994 (60): 1 – 22.

113. Kim, H. J., and D. Siegmund. The Likelihood Ratio Test for a Change – Point in Simple Linear Regression [J]. Biometrika. 1989 (76): 409 – 423.

114. Kim, L – M., and G. S. Maddala. Multiple Structural Breaks and Unit Roots in the Nominal and Real Exchange Rates. Unpublished manuscript. Department of Economics [M]. The University of Fiorida, 1991.

115. Koski, T. Hidden Markov Models for Bioinformatics [M]. Dordrecht: Kluwer Academic Publishers, 2001.

116. Lam, Pok – sang. The Hamilton Model with a General Autoregressive Component: Estimation and Comparison with Other Models of Economic Time Series [J]. Journal of Monetary Economics. 1990 (26): 409 – 432.

117. Lange, K. A quasi – Newton acceleration of the EM algorithm [J]. Statistica Sinica. 1995 (5): 1 – 18.

118. Lange, K. and Boehnke, M. Extensions to pedigree analysis V. Optimal calculation of Mendelian likelihoods [J]. Hum. Hered. 1983 (33): 291 – 301.

119. Lange, K. Mathematical and Statistical Methods for Genetic Analysis, second edition [M]. New York: Springer, 2002.

120. Lange, K. Optimization [M]. New York: Springer, 2004.

121. Le, N. D., Leroux, B. G. and Puterman, M. L. Reader reaction: Exact likelihood evaluation in a Markov mixture model for time series of seizure counts [J]. Biometrics. 1992 (48): 317 -323.

122. Leisch, F. FlexMix: A general framework for finite mixture models and latent class regression in R [J]. J. Statistical Software 11. 2004.

123. Leroux, B. G. and Puterman, M. L. Maximum - penalized - likelihood estimation for independent and Markov - dependent mixture models [J]. Biometrics. 1992 (48): 545 -558.

124. Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition [J]. Bell System Tech. J. 1983 (62): 1035 -1074.

125. Lindsey, J. K. R package "Repeated" . http: //popgen. unimaas. nl/~jlindsey/rcode. html, 27 -7 -2008.

126. Lindsey, J. K. Statistical Analysis of Stochastic Processes in Time [M]. Cambridge: Cambridge University Press, 2004.

127. Linhart, H. and Zucchini, W. Model Selection [M]. New York: Wiley, 1986.

128. Little, R. J. A. and Rubin, D. B. Statistical Analysis with Missing Data (second edition) [M]. New York: Wiley, 2002.

129. Lloyd, E. H. Handbook of Applicable Mathematics, Vol. 2: Probability [M]. New York: Wiley, 1980.

130. Lystig, T. C. and Hughes, J. P. Exact computation of the observed information matrix for hidden Markov models [J]. J. Comp. Graphical Statist. 2002 (11): 678 -689.

131. MacDonald, I. L. and Raubenheimer, D. Hidden Markov models and animal behaviour [J]. Biometrical J. 1995 (37): 701 -712.

132. MacDonald, I. L. and Zucchini, W. Hidden Markov and Other Models for Discrete - valued

Time Series [M]. London: Chapman & Hall, 1997.

133. Mandlebrot, Benoit. The Variation of Certain Speculative Prices [J]. Journal of Business, 1963 (4): 394 - 419.

134. McCullagh, P. and Nelder, J. A. Generalized Linear Models, second edition [M]. London: Chapman & Hall, 1989.

135. McFarland, D. Animal Behaviour: Psychobiology, Ethology and Evolution, third edition [M]. Harlow: Longman Scientific and Technical, 1999.

136. McLachlan, G. J. and Krishnan, T. The EM Algorithm and Extensions [M]. New York: Wiley, 1997.

137. McLachlan, G. J. and Peel, D. Finite Mixture Models [M]. New York: Wiley, 2000.

138. Mira, A. Exuviae eating: a nitrogen meal? [J] J. Insect Physiol. 2000 (46): 605 - 610.

139. Munoz, W. P., Haines, L. M. and van Gelderen, C. J. An analysis of the maternity data of Edendale Hospital in Natal for the period 1970 - 1985. Part 1: Trends and seasonality [R]. Internal report, Edendale Hospital, 1987.

140. Neftci, S. N. Are Economic Time Series Asymmetric over the Business Cycle? [J] Journal of Political Economy. 1984, 92 (2), 306 - 328.

141. Nelson, Charles R., and G. William Schwert. Short - Term Interest Rates as Predictors of Infiation: On Testing the Hypothesis That the Real Interest Rate is Constant [J]. American Economic Review. 1977 (67): 478 - 486.

142. Newton, M. A., and Raftery, A. E. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion) [J]. J. Roy. Statist. 1994, Soc. B 56: 3 - 48.

143. Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. and Bessìeres, P. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models [J]. Nucleic Acids Res. 2002 (30): 1418 - 1426.

144. Omori, Y., Chib, S., Shephard, N. and Nakajima, J. Stochastic volatility with leverage: fast and efficient likelihood inference [J]. J. Econometrics. 2007 (140): 425 - 449.

145. Pearl, J. Causality: Models, Reasoning and Inference [M]. Cambridge: Cambridge University

Press, 2000.

146. Pegram, G. G. S. An autoregressive model for multilag Markov chains [J]. J. Appl. Prob. 1980 (17): 350 - 362.

147. Perron, Pierre. Testing for a Unit Root in a Time Series with a Changing Mean [J]. Journal of Business and Economic Sitatistics. 1990, 8 (2): 153 - 162.

148. Ploberger, W., W. Kramer, and K. Kontrus. A New Test for Structural Stability in the Linear Regression Model [J]. Journal of Econometrics. 1989 (40): 307 - 318.

149. Quandt, R. E. The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes [J]. Journal of the American Statistical Association. 1958 (53): 873 - 880.

150. Quandt, R. E. A New Approach to Estimating Switching Regressions [J]. Journal of the American Statistical Association. 1972 (67): 306 - 310.

151. Quandt, R. E. Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes [J]. Journal of the American Statistical Association. 1960 (55): 324 - 330.

152. R Development Core Team. R: A language and environment for statistical computing [J]. R Foundation for Statistical Computing. Vienna, Austria, 2008.

153. Raftery A E. A model for high - order Markov chains [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1985: 528 - 539.

154. Raftery, A. E. A new model for discrete - valued time series: autocorrelations and extensions [J]. Rassegna di Metodi Statistici ed Applicazioni, 1985 (3 - 4): 149 - 162.

155. Raftery, A. E. and Tavar'e, S. Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model [J]. Appl. Statist. 1994 (43): 179 - 199.

156. Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). // J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds.). Bayesian Statistics 8. Oxford: Oxford University Press, 2007: 1 - 45.

157. Raubenheimer, D. and Barton Browne, L. Developmental changes in the patterns of feeding in fourth - and fifth - instar Helicoverpa armigera caterpillars. Physiol [J]. Entomology. 2000

(25): 390 - 399.

158. Raubenheimer, D. and Bernays, E. A. Patterns of feeding in the polyphagous grasshopper Taeniopoda eques: a field study [J]. Anim. Behav. 1993 (45): 153 - 167.

159. Richardson, S. and Green, P. J. On Bayesian analysis of mixtures with an unknown number of components (with discussion) [J]. J. Roy. Statist. 1997, Soc. B 59: 731 - 792.

160. Robert, C. P., Ryden T., Titterington D. M. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000, 62 (1): 57 - 75.

161. Robert, C. P. and Casella, G. Monte Carlo Statistical Methods [M]. New York: Springer, 1999.

162. Robert, C. P. and Titterington, D. M. Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation [J]. Statist. and Computing. 1998 (8): 145 - 158.

163. Rose, Andrew K. Is the Real Interest Rate Stable? [J] Journal of Finance. 1988 (43): 1095 - 1112.

164. Rosenblatt, M. Remarks on a multivariate transformation [J]. Ann. Math. Statist. 1952 (23): 470 - 472.

165. Rossi, A. and Gallo, G. M. Volatility estimation via hidden Markov models [J]. J. Empirical Finance. 2006 (13): 203 - 230.

166. Rydén, T., Teräsvirta, T., Åsbrink, S. Stylized facts of daily return series and the hidden Markov model [J]. Journal of applied econometrics, 1998, 13 (3): 217 - 244.

167. Schilling, W. A frequency distribution represented as the sum of two Poisson distributions [J]. J. Amer. Statist. Assoc. 1947 (42): 407 - 424.

168. Schimert, J. A high order hidden Markov model [D]. Ph. D. dissertation, University of Washington, 1992.

169. Scholz, F. W. Maximum likelihood estimation. // S. Kotz, N. Balakrishnan, C. B. Read, B. Vidakovic and N. L. Johnson (eds.). Encyclopedia of Statistical Sciences (second edition). Hoboken: Wiley, 2006: 4629 - 4639.

170. Sclove, S. L. Time - Series Segmentation: A Model and a Method [J]. Information Sciences. 1983 (29): 7 -25.

171. Scott, D. W. Multivariate Density Estimation: Theory, Practice and Visualization [M]. New York: Wiley, 1992.

172. Scott, S. L. Bayesian methods for hidden Markov models: Recursive computing in the 21st century [J]. J. Amer. Statist. Assoc. 2002 (97): 337 -351.

173. Scott, S. L., James, G. M. and Sugar, C. A. Hidden Markov models for longitudinal comparisons [J]. J. Amer. Statist. Assoc. 2005 (100): 359 -369.

174. Shephard, N. G. Statistical aspects of ARCH and stochastic volatility. //D. R. Cox, D. V. Hinkley and O. E. Barndorff - Nielsen (eds.). Time Series Models: In econometrics, finance and other fields. London: Chapman & Hall, 1996: 1 -67.

175. Sibly, R. M. and McFarland, D. On the fitness of behaviour sequences [J]. American Naturalist. 1976 (110): 601 -617.

176. Silverman, B. W. Density Estimation for Statistics and Data Analysis [M]. London: Chapman & Hall, 1986.

177. Silverman, B. W. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with Discussion) [J]. J. Roy. Statist. Soc. B 1985 (47): 1 -52.

178. Simpson, S. J. and Raubenheimer, D. The central role of the haemolymph in the regulation of nutrient intake in insects [J]. Physiol. Entomology. 1993 (18): 395 -403.

179. Simpson, S. J. The pattern of feeding. // R. F. Chapman and T. Joern (eds.). A Biology of Grasshoppers. New York: Wiley, 1990: 73 - 103.

180. Singh, G. B. Statistical Modeling of DNA Sequences and Patterns. // S. A. Krawetz and D. D. Womble (eds.). Introduction to Bioinformatics: A Theoretical and Practical Approach. Totowa: Humana Press, 2003: 357 -373.

181. Smyth, P., Heckerman, D. and Jordan, M. I. Probabilistic independence networks for hidden Markov probability models [J]. Neural Computation. 1997 (9): 227 -269.

182. Speed, T. P. Terence' s stuff: my favourite algorithm [J]. IMS Bulletin. 2008, 37 (9): 14.

183. Spreij, P. On the Markov property of a finite hidden Markov chain [J]. Statist. Prob. Letters. 2001 (52): 279 - 288.

184. Suster, M. L., Martin, J. R., Sung, C. and Robinow, S. Targeted expression of tetanus toxin reveals sets of neurons involved in larval locomotion in Drosophila [J]. J. Neurobiology. 2003 (55): 233 - 246.

185. Timmermann, A. Moments of Markov switching models [J]. J. Econometrics. 2000 (96): 75 - 111.

186. Titterington, D. M., Smith, A. F. M. and Makov, U. E. Statistical Analysis of Finite Mixture Distributions [M]. New York: Wiley, 1985.

187. Toates, F. Motivational Systems [M]. Cambridge: Cambridge University Press, 1986.

188. Turner, Christopher M., Richard Startz, and Charles R. Nelson. A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market [J]. Journal of Financial Economics. 1989 (25): 3 - 22.

189. Turner, R. Direct maximization of the likelihood of a hidden Markov model. Computat. Statist. & Data Analysis [J]. 2008 (52): 4147 - 4160.

190. van Belle, G. Statistical Rules of Thumb [M]. New York: Wiley, 2002.

191. Visser, I., Raijmakers, M. E. J. and Molenaar, P. C. M. Fitting hidden Markov models to psychological data [J]. Scientific Programming. 2002 (10): 185 - 199.

192. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. Inform. Th. 1967 (13): 260 - 269.

193. Walsh, Carl E. Three Questions Concerning Nominal and Real Interest Rates [J]. Economic Review, Federal Reserve Bank of San Franscisco, no. 4, 1987: 5 - 20.

194. Wasserman, L. Bayesian model selection and model averaging [J]. J. Math. Psychology. 2000 (44): 92 - 107.

195. Wecker, W. E. Predicting the Turning Points of a Time Series [J]. Journal of Business. 1979, 52 (1): 35 - 50.

196. Weisberg, S. Applied Linear Regression, second edition [M]. New York: Wiley, 1985.

197. Welch, L. R. Hidden Markov models and the Baum - Welch algorithm. IEEE Inform. Soc. Newsl. 2003 (53): 1, 10 - 13.

198. Whitaker, L. On the Poisson law of small numbers. Biometrika [J]. 1914 (10): 36 - 71.

199. Wittmann, B. K., Rurak, D. W. and Taylor, S. Real - time ultrasound observation of breathing and body movements in foetal lambs from 55 days gestation to term [C]. Abstract presented at the XI Annual Conference, Society for the Study of Foetal Physiology, Oxford, 1984.

200. Yu, J. On leverage in a stochastic volatility model [J]. J. Econometrics. 2005 (127): 165 - 178.

201. Zeger, S. L. and Qaqish, B. Markov regression models for time series: a quasi - likelihood approach [J]. Biometrics. 1988 (44): 1019 - 1031.

202. Zucchini, W. An introduction to model selection [J]. J. Math. Psychology. 2000 (44): 41 - 61.

203. Zucchini, W. and Guttorp, P. A hidden Markov model for space - time precipitation [J]. Water Resour. Res. 1991 (27): 1917 - 1923.

204. Zucchini, W. and MacDonald, I. L. Hidden Markov time series models: some computational issues. // S. Weisberg (ed.). Computing Science and Statistics 30. Interface Foundation of North America, Inc., Fairfax Station, VA, 1998: 157 - 163.

205. Zucchini, W. and MacDonald, I. L. Illustrations of the use of pseudoresiduals in assessing the fit of a model. // H. Friedl, A. Berghold, G. Kauermann (eds.). Statistical Modelling. Proceedings of the 14th International Workshop on Statistical Modelling, Graz, July 19 - 23, 1999: 409 - 416.

206. Zucchini, W., Raubenheimer, D. and MacDonald, I. L. Modeling time series of animal behavior by means of a latent - state model with feedback [J]. Biometrics. 2008 (64): 807 - 815.

207. Zucchini, Walter and Iain L. MacDonald. Hidden Markov models for time series: an introduction using R [M]. Chapman & Hall, 2009.

[General Information]
书名=解密复兴科技：基于隐蔽马尔科夫模型的时序分析方法
作者=刘振亚，邓磊著
页数=169
SS号=13532697
DX号=
出版日期=2014.04
出版社=中国经济出版社